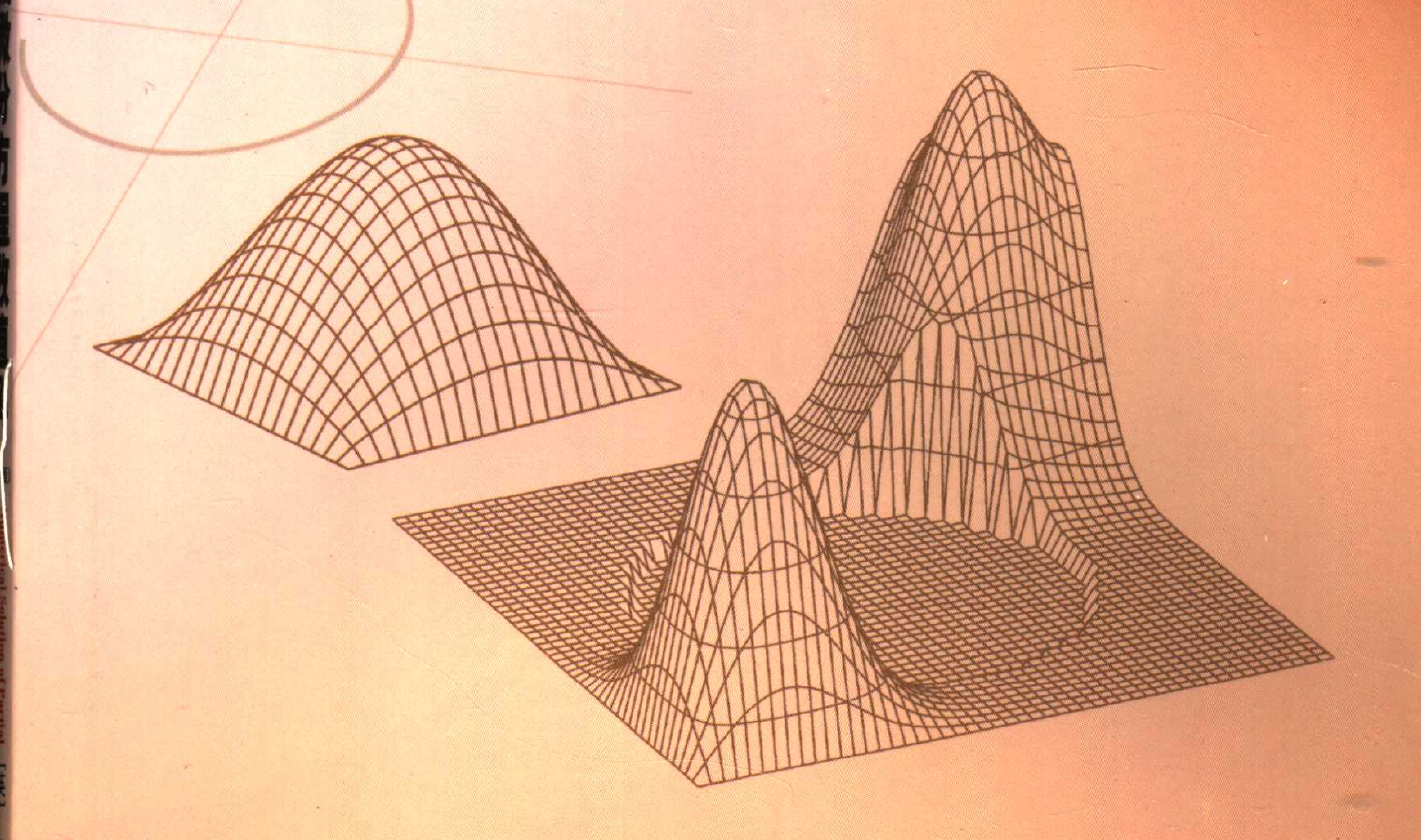
TURING 图灵数学·统计学丛书



Numerical Solution of Partial Differential Equations

偏微分方程数值解

（第2版）

[英] K. W. Morton
D. F. Mayers 著

李治平 门大力 许现民 张硕 译



偏微分方程数值解（第2版）

Numerical Solution of Partial Differential Equations

这是一本备受推崇的有关偏微分方程数值技术的教科书，被国外多家知名大学指定为教材。

本书讲解了求解偏微分方程的标准数值方法和技术，并蕴涵了该领域的最新发展。书中透彻地分析了各种方法的性质，严格地讨论了稳定性问题，提供了各种层次的例题和习题。全书结构清晰有序，叙述言简意赅。是数学、工程学及计算机科学专业学生学习偏微分方程数值解法首选的入门教材。

K. W. Morton 牛津大学退休教授，曾任教于数值分析学术重镇牛津大学计算实验室。现为巴斯大学兼职教授。主要研究领域为有限差分、有限元和有限体方法。Morton有着丰富的教学经验，他在数值分析领域理论研究和实际应用中的成就也广为人知。他曾担任数值分析界最高荣誉奖Leslie Fox评委会主席。

D. F. Mayers 曾任职于牛津大学计算实验室，是已故数值分析先驱 Leslie Fox的长期合作者。除本书之外，他还著有广为采用的教材*An Introduction to Numerical Analysis*。

本书相关信息请访问：图灵网站 http://www.turingbook.com
读者/作者热线：（010）88593802
反馈/投稿/推荐信箱：contact@turingbook.com

上架建议 数学/计算数学



人民邮电出版社网址 www.ptpress.com.cn

ISBN7-115-14203-3/TP·5090
定价:29.00 元

TURING 图灵数学·统计学丛书

Numerical Solution of Partial Differential Equations

偏微分方程数值解

（第 2 版）

[英] K. W. Morton　D. F. Mayers 著
李治平　门大力　许现民　张硕　译



图书在版编目（CIP）数据

偏微分方程数值解 /（英）莫顿，（英）迈耶斯著；李治平等译.
—北京：人民邮电出版社，2006.1
（图灵数学·统计学丛书）

ISBN 7-115-14203-3

Ⅰ. 偏…　Ⅱ. ①莫…②迈…③李…　Ⅲ. 偏微分方程—数值计算　Ⅳ. 0241.82

中国版本图书馆 CIP 数据核字（2005）第 143448 号

内容提要

偏微分方程是构建科学、工程学和其他领域的数学模型的主要手段. 一般情况下，这些模型都需要用数值方法去求解. 本书提供了标准数值技术的简明介绍. 借助抛物线型、双曲线型和椭圆型方程的一些简单例子介绍了常用的有限差分方法、有限元方法、有限体方法、修正方程分析、辛积分格式、对流扩散问题、多重网格、共轭梯度法. 利用极大值原理、能量法和离散傅里叶分析清晰严格地处理了稳定性问题. 本书全面讨论了这些方法的性质，并附有典型的图像结果，提供了不同难度的例子和练习.

本书可作为数学、工程学及计算机科学专业本科教材，也可供工程技术人员和应用工作者参考.

图灵数学·统计学丛书
偏微分方程数值解（第 2 版）

◆ 著　[英] K.W.Morton　D.F.Mayers
译　李治平　门大力　许现民　张　硕
责任编辑　王丽萍

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街 14 号
邮编　100061　电子函件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京顺义振华印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：800×1000　1/16
印张：14.5
字数：304 千字　2006 年 1 月第 1 版
印数：1 – 4 000 册　2006 年 1 月北京第 1 次印刷
著作权合同登记号　图字：01-2005-6096 号
ISBN 7-115-14203-3/TP·5090

定价：29.00 元

读者服务热线：(010)88593801　印装质量热线：(010)67129223

译 者 简 介

李治平 北京大学教授、博士生导师. 现任北京大学数学科学学院科学与工程计算系副系主任; 北京计算数学学会理事长, 中国计算数学学会常务理事; 《数值计算与计算机应用》和《高等学校计算数学学报》编委.

1982 年毕业于西安交通大学, 获理学学士学位; 1982～ 1987 年就读于北京大学, 获理学硕士、博士学位. 1988～1990 年在英国 Heriot-Watt 大学做博士后研究, 任 Research Associate ; 1991～1993 年在英国 Brunel 大学做博士后研究, 任 Research Fellow. 曾先后在英国伦敦中央工业学院、美国宾州州立大学、普林斯顿大学、普林斯顿高等研究所、奥地利维也纳工业大学、瑞士苏黎世大学、香港浸会大学作访问学者. 在偏微分方程数值解方面从事了多年的研究工作.

译 者 序

偏微分方程的数值方法不仅是计算数学专业的一门重要的课程，而且有越来越多的理工科各专业的学生要求学习这门课程. 目前，有关的教材已经很多. 然而，这本由 K.W. Morton 和 D.F. Mayers 为牛津大学数学专业本科生编写的《偏微分方程数值解》确实是一本很有特点的好教材. 它通过一些简单的模型问题重点介绍了有限差分方法，在离散范数下讨论了算法的相容性、稳定性和收敛性，并且介绍了修正方程分析等用于分析算法引起的耗散、色散等现象及分析算法稳定性的有力工具；利用极大值原理、能量法和离散傅里叶分析清晰严格地处理了稳定性问题；同时也简略地介绍了有限元方法、有限体积法等经典的具有一般性的方法，还介绍了多重网格法、共轭梯度法等一些实用的迭代算法，求解由偏微分方程离散化得到的线性代数方程组可使用这些算法；还介绍了如多辛格式这样的该领域的最新进展. 该书在讲述方法的同时，还注意介绍这些方法的发展历史、设计思想和理论依据，并给出了相当丰富的参考文献. 该书内容丰富，但其所需的数学基础知识却相对较少，所以很适合将其作为一门独立的课程或偏微分方程理论的辅助课程，较早地引入到数学本科以及理工科其他专业本科高年级和研究生的教学中.

在北京大学数学学院科学与工程计算系，原来为本科高年级学生开设了有限差分方法和有限元方法两门各一个学期的课程. 然而近几年，为了适应学科发展和课程设置等方面的需要，将两者合并成了偏微分方程数值解这门一个学期的课程. 由于课时的减少，在教学内容的取舍，数学基础难易程度的掌握等方面可以说是矛盾重重. K. W. Morton 和 D.F. Mayers 编写的《偏微分方程数值解》这本教材为我们提供了很好的范例，很值得我们参考借鉴.

我们很高兴能有机会将该书推荐给国内的读者.

本书的翻译工作是由我和我的 3 名博士生合作完成的. 第 1 、 3 、 5 章由我自己翻译，其他章节由 3 名博士生翻译（第 2 章由门大力翻译，第 4 章由许现民翻译，第 6 、 7 章由张硕翻译），全书由我负责统校 [1].

由于译者无论是英文还是中文水平都有限，难免会有不妥之处，欢迎广大读者批评指正.

李治平

2005 年 11 月于燕园

[1] 出版者说明：本书由 LaTeX 排版，出版者感谢译者辛勤的工作，也感谢北京师范大学李勇教授对我们排版工作的帮助.

第一版序

本书源于我们两人在过去几年中分别给牛津大学数学系本科毕业班所讲的课程，它共包含 16 讲. 我们的同事建议把它作为偏微分方程入门理论的一个配套课程更早些面向学生开设，这促使我们编写了这本书. 另一方面，我们也在给硕士研究生讲授数学建模和数值分析这门课时使用了大致相同的内容. 基于这两方面的考虑，我们选择了适当的主题、难度和教授方式.

我们主要关注有限差分方法和它们在标准模型问题中的应用，这样在比较严格地处理如收敛性和稳定性等数学概念的同时，可以用简单的术语来描述方法. 在更高级的或者更偏重有限元方法的课程中，使用标准的索伯列夫范数会更自然和方便. 然而我们没有这样选择，仅仅使用了离散范数，确切地说是最大值范数和 l_2 范数. 我们有许多理由这么做. 首先，这自然与需要尽量少的背景知识 (包括分析、偏微分方程理论和计算等方面) 的目的相一致，因此本书可以作为本科生的早期课程适用于理工科和数学系的学生.

同等重要地是，这与我们分析椭圆型和抛物型问题的算法时广泛应用离散最大值原理，与我们处理离散能量方法与守恒律，研究有限区域上傅里叶波型等方面的做法相一致. 我们相信从纯粹的离散层面来论述所有这些概念有助于加强学生对这些重要数学工具的理解. 同时这也是非常实用的方法，并且有助于将差分格式理解为物理原理的直接模型：毕竟，所有的计算都在有限网格上进行，而且实际计算的稳定性等也是在离散层面来检测的. 进而，用一个时间步长上的阻尼和色散来解释差分格式作用于能在网格上表示的傅里叶波型所产生的效果比使用截断误差价值更大，这也例证了本书之所以采取目前做法的第二个理由.

尽管如此，为了正确地理解偏微分方程的数值方法，了解随网格参数 h 趋向于零的极限过程是至关重要的. 例如，如果 U^n 是在第 n 个时间层的离散逼近，并且其关于时间步长 Δt 的变化由 $U^{n+1} = C_h U^n$ 表示，许多学生觉得要区分当固定网格和固定 Δt 时 $n \to \infty$ 的极限过程，和当 $n\Delta t$ 固定而 h, $\Delta t \to 0$ 时 $n \to \infty$ 的极限过程非常困难. 这两个过程都有其实用方面的重要性：许多定常问题的离散方程求解常常仿照非定常问题的时间步迭代过程，前者与此过程有关；而理解后者在选择逼近非定常问题的方法以避免不稳定性时极其重要. 一致有界和一致收敛的概念是很关键的；当然，如果使用和 h 无关的范数会更容易处理. 但是 Palencia 和 Sanz-Serna[1] 的例子指出，基于离散范数可以建立

[1] Palencia, C. and Sanz-Serna, J.M.(1984), An extension of the Lax-Richtmyer theory, *Numer. Math.* **44**(2), 279-283.

一套严格的理论，而这正是我们所采用方法的基础．它也意味着定义例如适定性等概念时必须十分小心；虽然这里的处理略显笨拙，但是它在实用和教学法方面却带来了很大的好处．

主题顺序的细心安排体现了以上的思想．我们从处理抛物型问题开始，它在逼近和分析方面都是最简单的，同时也具有最广泛的应用．通过在扩散算子上引入附加的对流项，就很自然地引到双曲型问题的研究．在细致地讨论这两种情况后，我们才在第 5 章对发展问题严格地给出相容性、收敛性和稳定性的概念．最后两章分别讲述椭圆型问题的离散化及有限元方法引论和求解由此得到的代数方程组的迭代解法，后者和抛物型方程解之间的紧密联系使主题间的关联更加完整．在所有情况下，我们只给出了为数不多的几种算法，但每一种都给出了全面的分析，并且证实了它们的实用性．事实上，本书恪守了一个宗旨，就是所有的模型问题以及逼近它们的方法都很简单且具有一般性．

每章的结尾有难度各异的习题；它们完善、拓展或者只是例述了正文中的内容．它们基本上都是分析类的习题，所以，全书可以作为一门纯理论课程的教材．尽管这样，因为数值分析有很实用的目标，所以在正文中都有各个方法的数值演示；依照这些例子可以容易地教学生构造进一步的数值试验．计算工具和实践的快速发展让我们相信，这种可扩充的方法比给出显式的实际练习更有利．

我们用两种方式引用相关文献．在正文中引入关键的概念以及其与特定的原始论著相关联时，在注脚中给出了完整的参考资料，前言就已经采用了这种形式．另外，在每章的结束处包含简短的一节名为“文献注记与推荐读物”．这两种方式所提供的参考文献都不是全面的，但是，它们足够帮助有兴趣的学生进一步学习相关内容．当然 Richtmyer 和 Morton(1967) 处理发展问题的方法给予了我们很大的启示和影响，本书可以作为他们那本书的引论和补充．

感谢阅读本书早期版本的同事和他们给出的许多建议 (特别是 Endre Süle 有益的评注)．同时也感谢检验习题的学生们的帮助．在本书很长的成形过程中，我们的秘书 Jane Ellory 和 Joan Himpson 细致耐心的工作对本书的完成同样至关重要．

第二版序

本书第一版出版后的10年中，偏微分方程数值解法在许多方面都有了发展．但是当我们询问在第二版中主要应该做何种更改时，得到的大多数反馈是不应该更改本书的要点或者说只做非实质的改变．所以我们要求自己加入不超过 10%~20% 的新内容并且很少删减前版内容：结果本书增加了大约 23% 的内容．

无论从理论还是从处理实际问题工具的角度来说，有限差分方法依然是介绍偏微分方程数值解法的出发点，所以它们仍然是本书的核心．当然椭圆型方程领域有限元方法占主导地位，而在逼近许多双曲型问题时有限体积法更占优势．进而，有限体积法是两种主要方法间有用的桥梁．因此我们在第 4 章加入了一节讲述此主题，而且我们可以用这个方法重新解释标准的差分格式，例如 Lax-Wendroff 方法和盒式格式，然后例示如何简单地把它们推广到非一致网格．另外，在第 6 章介绍有限元方法后，新加入了关于对流扩散问题的一节：它涵盖了有限差分和有限元方法，并且引入了 Petrov-Galerkin 方法．

有限差分方法的理论框架完整建立已久，所以不需要做太多修改．可是在过去几年中逼近常微分方程和偏微分方程的方法之间发生了许多相互的影响，因而我们也从几方面体现了这一点．首先，人们将辛方法应用于哈密顿常微分方程系统以及推广应用于偏微分方程的兴趣日益增长，因此我们在第 4 章用一节讨论了这个主题，并且用这个思想去分析了用于逼近波动方程的交错蛙跳格式．更一般地，在第 5 章中对线法的重新关注让我们完全改写了关于稳定性分析的能量方法一节，这不仅使特例分析而且使整体一致性得到了重要的改善．在这一章还加入了介绍修正方程分析一节．虽然这是一个 30 年前引入的技术，但是这种方法更好地解释了盒式格式，让人们开始重新评估它的价值．进而，它在常微分方程逼近中的应用不但加强了它在分析方面的地位，而且其重要性也得到了更广泛的认可．

在偏微分方程数值解领域中，我们所描述的方法在实际应用中发生了巨大的变化．为了在新版中依然秉承恰当地介绍和反映实用方法的原则，本书做了一些改进．特别地，对大规模代数方程组系统迭代解法的处理有了极大的改进．这导致了在依赖时间的问题中更加大量地使用隐式格式，在椭圆型问题的有限元模拟中用迭代法代替直接法，以及处理这两种类型问题的方法之间更密切的相互影响．因此第 7 章的重点已经改变并新增了两节，分别介绍十分重要的多重网格方法和共轭梯度法，它们也正是实际计算中发生如此大变化的主要原因．

我们曾经考虑在本书中包含一定数量的 Matlab 的程序来演示一些主要方法．但是考

虑到最近 10 年个人计算机及其软件的飞快发展，我们意识到这样的材料会很快过时，因此本书的这一部分没有改变. 我们像第一版一样处理文献资料和文献注记，不过在本书末尾处把它们都汇集到了参考文献目录中.

每章结尾的练习都有 LaTeX文件格式的答案. 若教授这方面的课程，只要发送电子函件到 solutions@cambridge.org 就可以获得.

我们非常感谢指出本书第一版中错误的读者们，同时希望我们在第二版中完全更正了这些错误并且没有再引入新的错误. 再次感谢我们的同事对新版本的阅读和建议.

目　录

第 1 章　引言　1

第 2 章　一维抛物型方程　5
2.1 引论　5
2.2 模型问题　5
2.3 级数逼近　6
2.4 模型问题的显式格式　7
2.5 差分格式和截断误差　10
2.6 显式格式的收敛性　12
2.7 误差的傅里叶分析　15
2.8 隐式方法　17
2.9 Thomas 算法　18
2.10 加权平均和 θ-方法　20
2.11 最大值原理和 $\mu(1-\theta)\leqslant\frac{1}{2}$ 时的收敛性　25
2.12 三时间层格式　30
2.13 更一般的边界条件　30
2.14 热量守恒性质　34
2.15 更一般的线性问题　36
2.16 极坐标　40
2.17 非线性问题　42
文献注记与推荐读物　44
习题　45

第 3 章　二维和三维抛物型方程　49
3.1 盒形区域上的显式方法　49
3.2 二维 ADI 方法 (交替方向迭代法)　50
3.3 三维 ADI 和 LOD 方法　55
3.4 曲线边界　56
3.5 应用于一般抛物型问题　61
文献注记与推荐读物　65
习题　65

第 4 章　一维双曲型方程　69

4.1 特征线方法 69
4.2 CFL 条件 71
4.3 迎风格式的误差分析 75
4.4 迎风格式的傅里叶分析 77
4.5 Lax-Wendroff 格式 80
4.6 守恒律的 Lax-Wendroff 方法 83
4.7 有限体积格式 89
4.8 盒式格式 93
4.9 蛙跳格式 99
4.10 哈密顿系统与辛积分格式 103
4.11 相误差和振幅误差的比较 109
4.12 边界条件与守恒性质 110
4.13 高维情形 115
文献注记与推荐读物 117
习题 117

第 5 章 相容性、收敛性和稳定性 **121**
5.1 问题的定义 121
5.2 有限差分的网格与范数 122
5.3 有限差分逼近 124
5.4 相容性、精度的阶和收敛性 125
5.5 稳定性与 Lax 等价定理 126
5.6 稳定性条件的计算 128
5.7 实用的（严厉的或强的）稳定性 133
5.8 修正方程分析 135
5.9 守恒律与能量法分析 141
5.10 理论综述 148
文献注记与推荐读物 151
习题 151

第 6 章 二维线性二阶椭圆型方程 **155**
6.1 一个模型问题 155
6.2 模型问题的误差分析 155
6.3 一般的扩散问题 157
6.4 曲线边界上的边值条件 159
6.5 利用最大值原理的误差分析 162

6.6 渐近误差分析 170
6.7 变分形式和有限元方法 174
6.8 对流扩散问题 179
6.9 一个例子 182
文献注记与推荐读物 184
习题 185

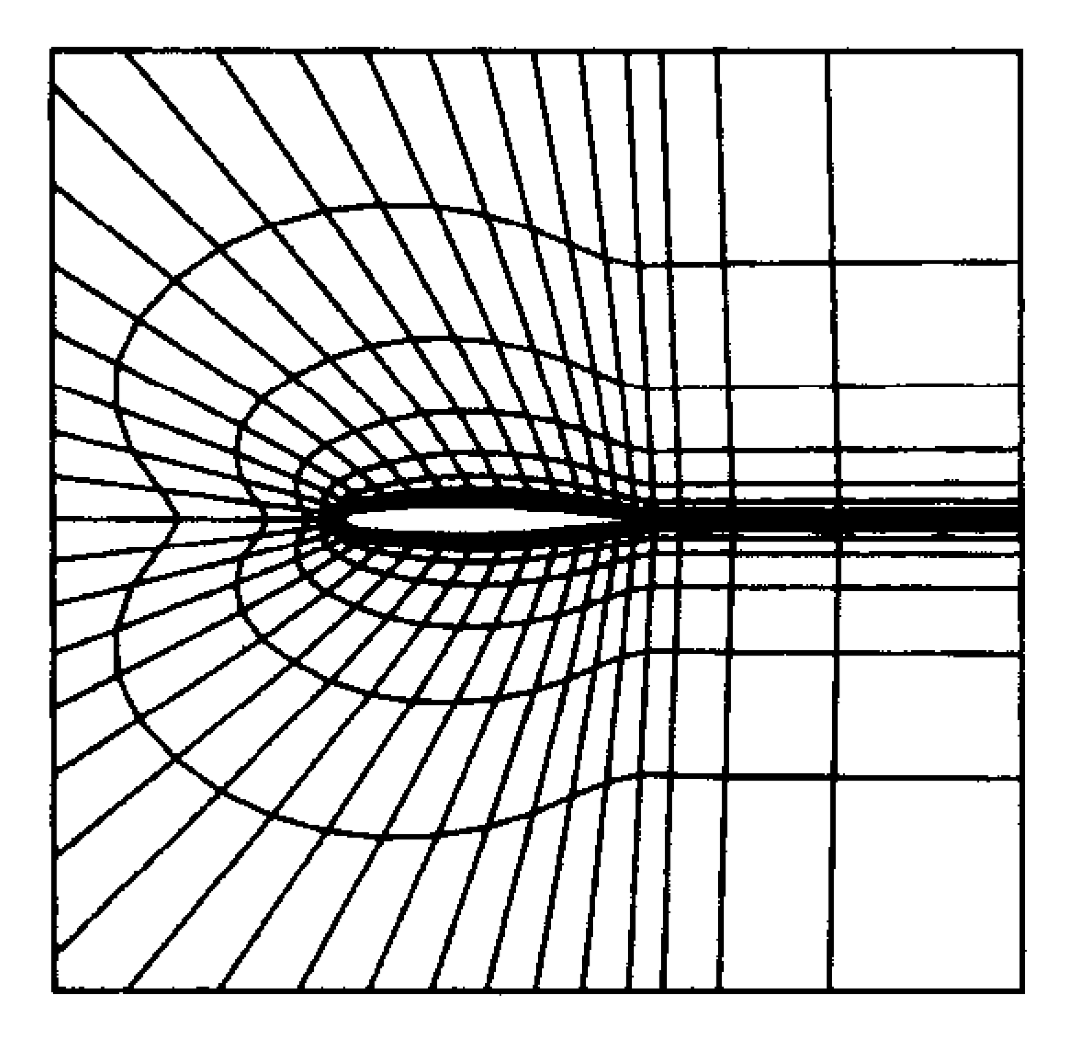
第 7 章 线性代数方程组的迭代求解 **189**
7.1 显式基本迭代格式 190
7.2 迭代法的矩阵形式及其收敛性 192
7.3 收敛性的傅里叶分析 195
7.4 应用于一个例子 199
7.5 推广及相关的迭代法 201
7.6 多重网格法 202
7.7 共轭梯度法 206
7.8 数值例子：几个对比 209
文献注记与推荐读物 210
习题 211

其他参考文献 **213**

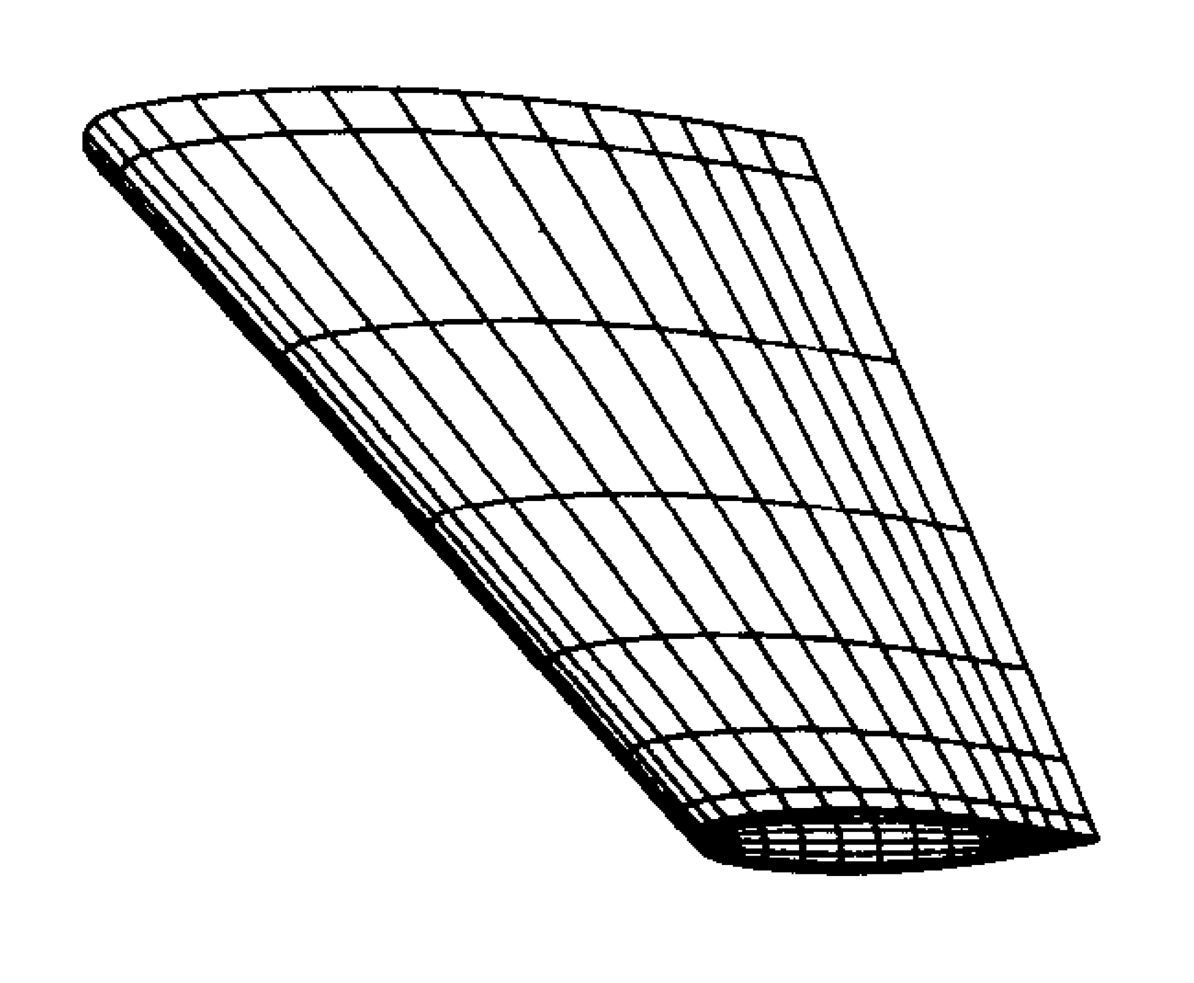
第 1 章　引言

偏微分方程（ PDE ）是众多描述物理、化学和生物现象的数学模型的基础，而且其最新的应用已经扩展到经济、金融预测、图像处理及其他领域. 要通过相应的PDE 模型研究这些现象，我们通常需要结合一些对简单特例的分析，采用数值方法求得这些方程的近似解；而在一些最新的应用中数值模型已经几乎是独当一面了.

我们来考虑如图 1-1 所示的机翼设计问题. 当然，还有许多可供我们选择的好例子，例如天气预报、污染物的扩散、喷气发动机或内燃机的设计、核反应堆的安全性、石油的勘探与开采等等.



（a）一种典型的（无粘）环绕机翼截面的计算网格



（b）机翼表面的相应网格

图 1-1　计算网格

在*定常* (steady) 飞行状态中，机翼两个重要的设计指标是由空气流过机翼而产生的升力和阻力. 我们为一个设计方案计算这些量时，根据*边界层* (boundary layer) 理论，在很好的近似意义下，在靠近机翼表面有一个很薄的边界层，层内粘力起着重要的作用，而在边界层之外，可以假设气流是*无粘的* (inviscid). 因此，假设机翼是局部平坦的，则机翼附近的气体流动可由以下模型方程描述

$$u\frac{\partial u}{\partial x}-\nu\frac{\partial^2 u}{\partial y^2}=(1/\rho)\frac{\partial p}{\partial x}\ , \tag{1.1}$$

其中 u 是切空间坐标 x 方向的流动速度， y 是法向坐标， ν 是粘性， ρ 是密度， p 是压力；我们在该方程中忽略了法向速度. 这是一个典型的关于 u 的 *抛物型方程* (parabolic

equation)，其中 $(1/\rho)\partial p/\partial x$ 作为外力项.

在机翼边界层之外，仅限于二维横截面 (cross section) 上，我们可以假设气流是无粘的且其速度可以表示为 $(u_\infty + u, v)$，其中 u 和 v 与气流在无穷远处沿 x 方向的速度 u_∞ 相比是小量. 我们通常还可以假设气流是无旋的 (irrotational)，因此有

$$\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = 0\,; \tag{1.2a}$$

假设气流是等熵的 (homentropic)，则结合质量和 x 方向动量的守恒律 (conservation laws) 且仅保留一阶小量，我们就得到了简化的模型

$$(1 - M_\infty^2)\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0\,, \tag{1.2b}$$

其中 M_∞ 是无穷远处的马赫数 (Mach number)， $M_\infty = u_\infty/a_\infty$，$a_\infty$ 是声速.

显然当气流是亚音速 (subsonic) 时，即 $M_\infty < 1$ 时，方程组 (1.2a, b) 与柯西 - 黎曼 (Cauchy-Riemann) 方程等价，是椭圆型方程组 (elliptic equations)．而对于超音速流 $M_\infty > 1$，该方程组等价于一维波动方程，为双曲型方程组 (hyperbolic equations)．我们也可以在 (1.2b) 上作用 $\partial/\partial x$，在 (1.2a) 上作用 $\partial/\partial y$ 并消去 v，从而得到与 Laplace 方程或二阶波动方程 (wave equation) 等价的方程组.

因此，我们仅从这一个问题中就精练出了三种基本类型的偏微分方程，当然，从我们前面提到的其他问题中也同样可以得到这三种基本类型的偏微分方程. 由 PDE 理论，我们知道这三种类型偏微分方程的分析性质、问题适定性 (well-posedness) 的提法、边界条件 (boundary condition) 的提法及其解的性质，所有一切都差别巨大. 对于这三种基本类型的偏微分方程的数值求解及其数值分析来说，情况也是如此.

在本书中，我们将只讨论这三类方程的模型问题，因为理解这些模型问题是研究更复杂的偏微分方程组的重要基础. 我们将讨论算法，主要是有限差分方法 (finite difference method) 及密切相关的有限体积法 (finite volume method)，这些方法可以应用于解决更实际、更复杂的问题，但只有在简单的情况下才能得到深入透彻的分析结果. 对于模型问题，我们将能够建立当有限差分网格加密时算法的稳定性 (stability)、收敛性 (convergence) 等严格的分析理论. 同样，我们能够细致地研究求解由差分方法得到的代数方程组的迭代法 (iterative method) 的收敛速度 (speed of convergence)，所得到的结果可以广泛地应用于无法进行精确分析的实际问题中去.

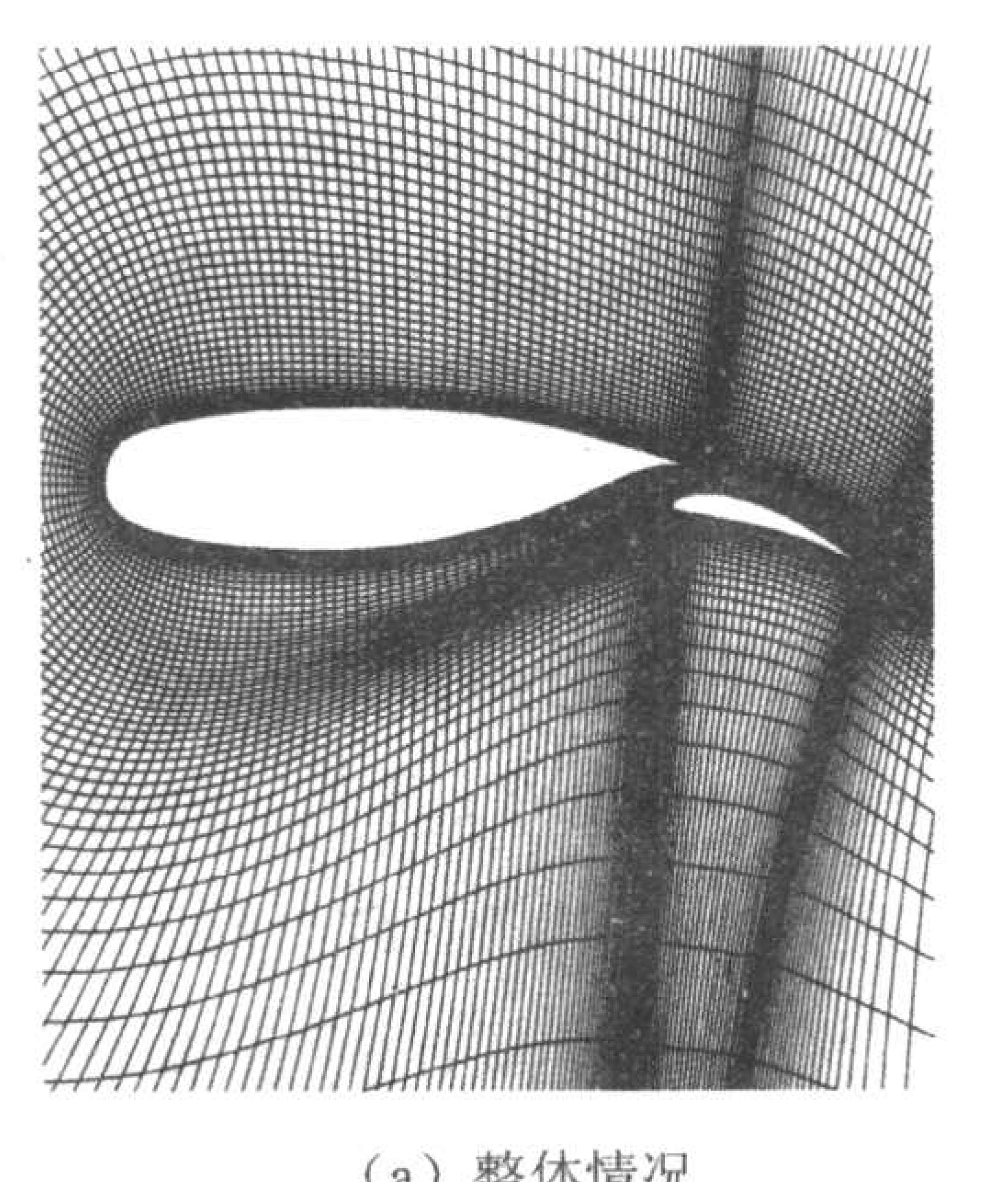
尽管本书的重点放在分别单独讨论某一类型的方程上，我们必须指出的是在许多实际问题中这些类型的方程会同时出现在一个方程组中. 有着广泛应用的欧拉—泊松 (Euler-Poisson) 方程组就是这样一个例子：在空间变量为二维，时间为 t 时，该方程组是关于两个速度分量和前面所提到过的压力这 3 个变量的偏微分方程组；若使用更紧凑的

记号 (例如用 ∂_t 代替 ∂/∂_t)，则欧拉—泊松方程组可以表示为以下形式

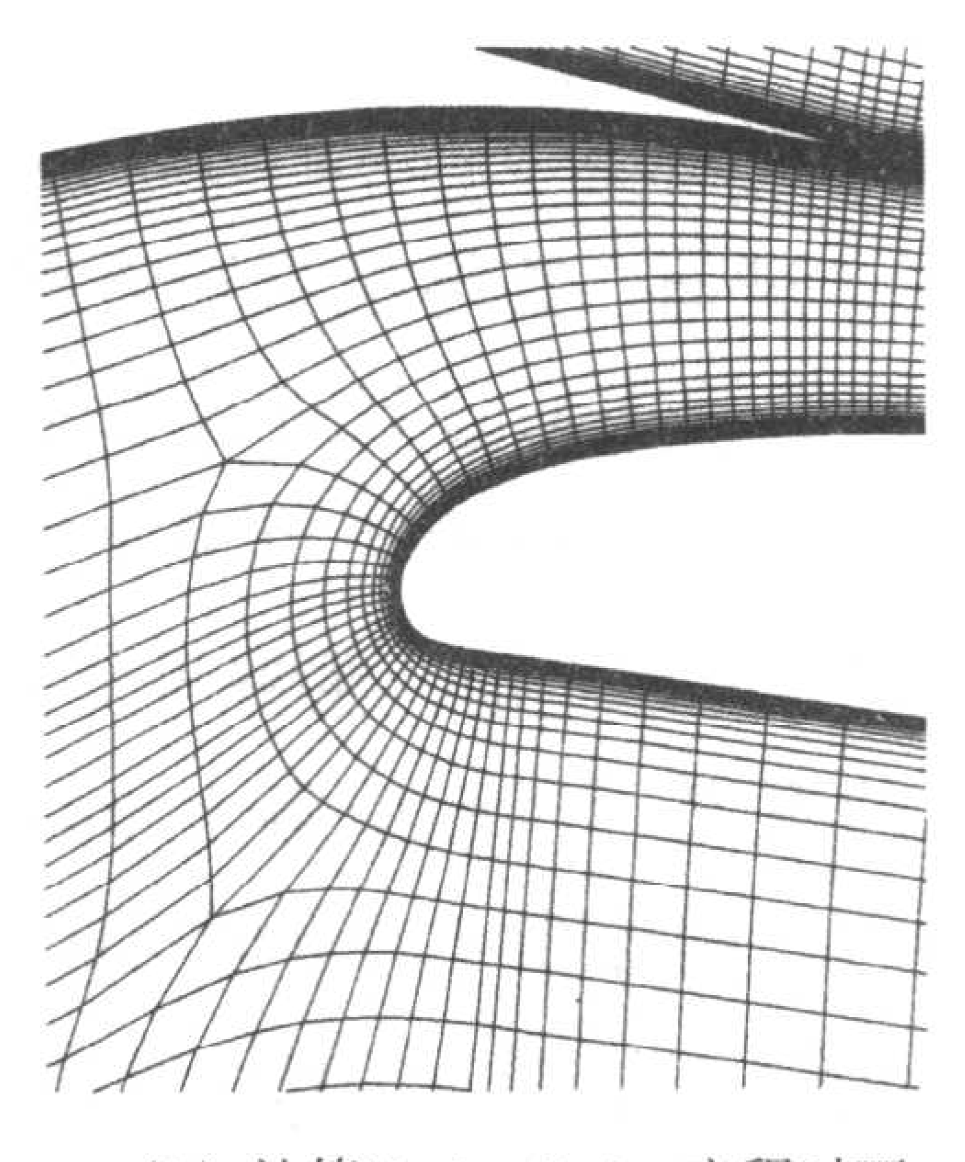
$$\partial_t u + u\partial_x u + v\partial_y u + \partial_x p = 0$$
$$\partial_t v + u\partial_x v + v\partial_y v + \partial_y p = 0$$
$$\partial_x^2 p + \partial_y^2 p = 0. \tag{1.3}$$

求解该方程组时需要结合两种非常不同的方法：对于最后一个关于 p 的椭圆型方程，需要采用第 6 、 7 两章讲述求解大型代数方程组的方法，其解又为前两个双曲型方程组提供了外力项；而对于双曲型方程组一般可采用第 $2\sim5$ 章讲述的方法，在时间方向上以逐步推进的方式求解. 这类模型通常源于流速远低于空气动力学中气流速度的问题，例如多孔介质中的流体运动 (如地下水的流动). 这两种类型的方法需要紧密地耦合起来才能有效地发挥作用.

回到前面机翼设计的例子，我们不妨提一下实际当中会出现的一些复杂情况. 对于民用飞机，最主要的是考虑其在设计速度下定常飞行时的性能；然而，对于军用飞机，操控性也是很重要的，也就是说要考虑气流是非定常的，即方程包含时间变量的情况. 这时，即便是亚音速流，相应于 (1.2a, b) 的方程组也是双曲型的 (一个时间变量和两个空间变量)，该方程组与欧拉—泊松方程组 (1.3) 类似但更为复杂. 同时还必须考虑更大的几何复杂度：特别是在计算机翼尖端和机翼与机体的结合部附近的气流时，必须考虑三维的机翼；而在着陆和起飞阶段，为在低速时提供更大的升力，飞机的襟翼会展开，这时机翼的横截面就会如图 1-2 所示.



(a) 整体情况



(b) 计算Navier-Stokes方程时可能会用到的网格局部细节

图 1-2　一种典型的复合型机翼 (DRA, Farnborough 提供)

另外，我们一直假设流动是光滑的，但在实际中我们需要研究诸如激波、涡片、湍流及其相互间的作用. 我们将要研究建立的方法可以应用于模拟所有这些现象，但其所涉及的内容远远超出了本书的范围. 目前，工业界可以做到的事包括，近似求解绕整架飞机流

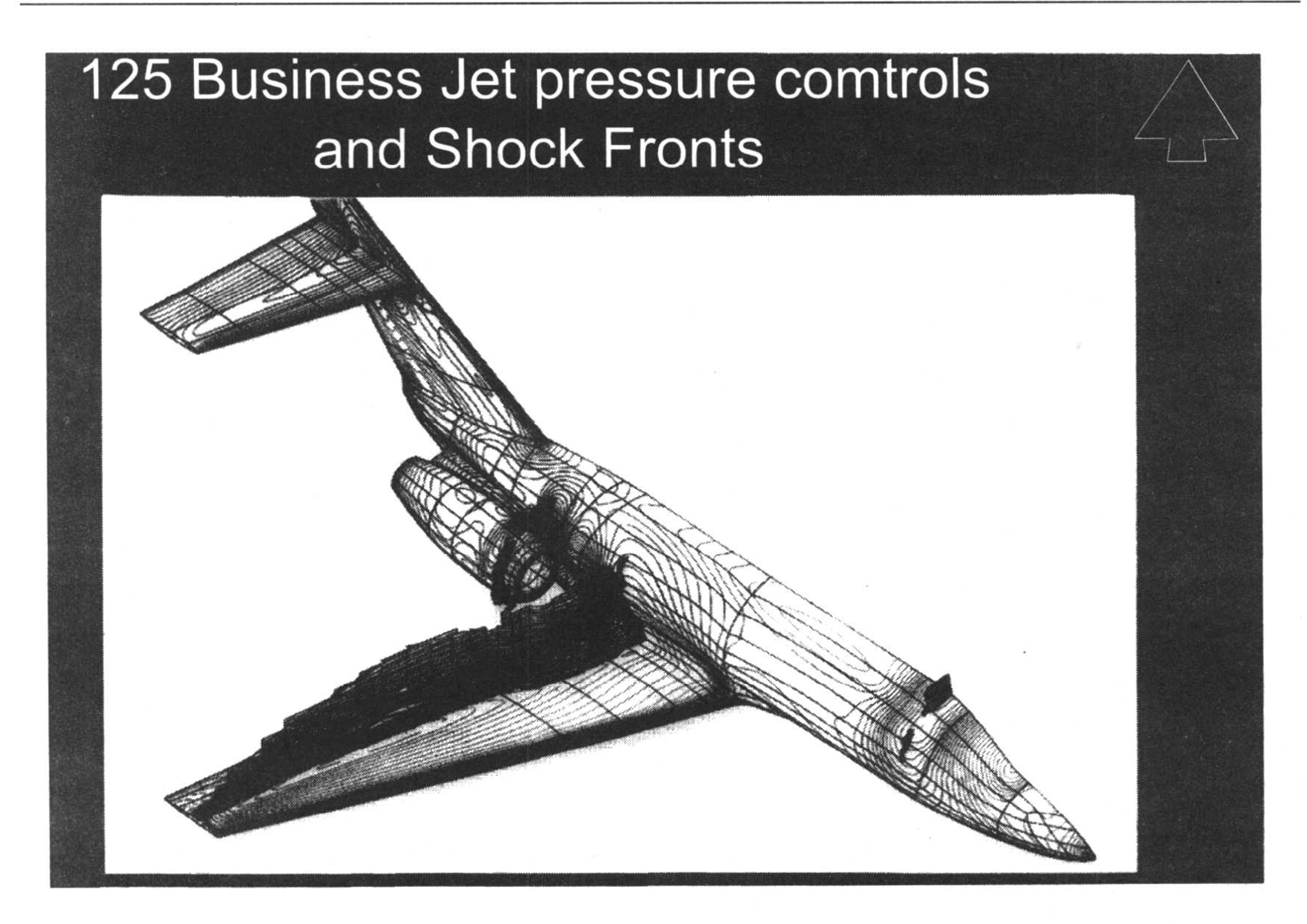


图 1-3 巡航条件下整架飞机的压力等高线和部分网格 (British Aerospace 提供)

动的 非定常粘性流 (unsteady viscous flow) 的平均 雷诺数 (Reynolds-averaged) Navier-Stokes 方程，如图 1-3 所示，进而，其终极目标是将这些对流体运动的预测能力融合到飞机整体设计的环节中去：我们希望通过设计飞机的形状使气流按给定的方式绕飞机运动，而不是仅仅满足于计算绕给定形状的飞机运动的气流.

最后，在结束引言这章之前，在符号方面有几点说明提请读者注意. 我们通常用符号 $\approx$ 表示在数值计算的意义下 "近似等于 (approximately equal to) "，而符号 $\sim$ 则表示 "渐近于 (asymptotic to) "，例如当 $t \to 0$ 时 $f(t) \sim t^2$ 表示 $t \to 0$ 时 $t^{-2}[f(t) - t^2] \to 0$，这也可以记为 $f(t) = t^2 + o(t^2)$； $f(t) = O(t^2)$ 则表示 $t^{-2}f(t)$ 当 $t \to 0$ 时是有界的. 我们通常用符号 $:=$ 表示等式左端是由等式右端 定义的(defined). 向量一般用粗体字表示.

第 2 章　一维抛物型方程

2.1　引论

本章讨论含有一个空间变量和一个时间变量 t 的抛物型方程 (parabolic equation) 的数值解法. 我们从最简单的均匀介质中的热传导模型出发. 对这一模型问题 (model problem) 可以直接采用显式差分格式, 用最大值原理或者傅里叶分析 (Fourier analysis) 的方法也很容易给出它的误差分析. 然而我们将会看到, 如果不严格限制时间步长的大小, 数值解将是不稳定的. 所以, 我们需要进一步考虑其他更为精确的数值方法来避免这种限制. 时间方向所需离散步数的减少所带来的好处远远超过了由此产生的额外的数值计算复杂性. 然后, 我们把这些方法推广到具有更一般边界条件的问题, 以及更一般的线性 (linear) 抛物型方程. 最后, 讨论更复杂的非线性方程 (nonlinear equation) 的解法.

2.2　模型问题

科学和工程中的许多问题常用关于未知函量 $u(x,t)$ 的某种特殊形式的线性抛物型方程来描述

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(b(x,t)\frac{\partial u}{\partial x}\right)+c(x,t)u+d(x,t), \tag{2.1}$$

其中 b 是严格正的. 在 $t=0$ 时, 我们取如下形式的 *初始条件* (initial condition)

$$u(x,0)=u^0(x), \tag{2.2}$$

其中 $u^0(x)$ 是给定函数. 问题的解在 $t>0$ 和 x 属于一个开区域 R 时满足 (2.1). R 一般是整个实轴、半实轴 $x>0$ 或者一个如 $(0,1)$ 的开区间. 在后两种情形, 要求解在 R 的闭包上有定义并且满足一定的 *边界条件* (boundary condition). 我们将假定边界条件关于 u 或者它的一阶空间导数 $\partial u/\partial x$ 是线性的, 或者二者都是线性的. 如果 $x=0$ 是左边界, 边界条件形式如下

$$\alpha_0(t)u+\alpha_1(t)\frac{\partial u}{\partial x}=\alpha_2(t), \tag{2.3}$$

其中

$$\alpha_0\geqslant 0,\ \alpha_1\leqslant 0\ 且\ \alpha_0-\alpha_1>0. \tag{2.4}$$

如果 $x=1$ 是右边界，则需要如下形式的边界条件

$$\beta_0(t)u+\beta_1(t)\frac{\partial u}{\partial x}=\beta_2(t), \tag{2.5}$$

其中

$$\beta_0\geqslant 0,\ \beta_1\geqslant 0\ 且\ \beta_0+\beta_1>0. \tag{2.6}$$

关于系数 α 和 β 的限制条件在后面的章节中会给出解释. 注意 α_1 和 β_1 之间符号的变化是因为 $\partial/\partial x$ 在右边界是外法向导数，而在 (2.3) 中它是内法向导数.

先考虑一个简单的模型问题，这个模型的方程描述了在有限区域内不随时间变化的均匀介质中无热源热流的运动. 假设给定齐次 *狄利克雷边界条件* (Dirichlet boundary condition), 也就是对所有的 t 方程的解在区域的两端等于零. 作无量纲变量替换后，问题变为：寻找定义在 $x\in[0,1]$ 和 $t\geqslant 0$ 的 $u(x,t)$ 满足

$$u_t=u_{xx},\qquad t>0,\ 0<x<1, \tag{2.7}$$

$$u(0,t)=u(1,t)=0,\qquad t>0, \tag{2.8}$$

$$u(x,0)=u^0(x),\qquad 0\leqslant x\leqslant 1. \tag{2.9}$$

在此，我们引入了常用的下标符号表示 *偏导数*(partial derivative).

2.3 级数逼近

这类微分方程可以用 *分离变量法* (separation of variable) 找出一些特解. 但是与有限差分方法不同的是分离变量法在应用中有很大的局限性，所以我们主要的着眼点还是有限差分方法. 尽管如此，它给出了可以用于比较的真解，同时导出了可自然地用于讨论有限差分方法稳定性的傅里叶分析方法.

寻找 $u(x,t)=f(x)g(t)$ 形式的特解，代入微分方程中得到

$$fg'=f''g,$$

即

$$g'/g=f''/f. \tag{2.10}$$

最后一个方程式中，左端独立于变量 x, 右端独立于变量 t, 因此两端必定均为常数. 记此常数为 $-k^2$, 直接求解两个分别关于函数 f 和 g 的简单的常微分方程，便得到解

$$u(x,t)=\mathrm{e}^{-k^2t}\sin kx.$$

从上式可以看出选取常数为 $-k^2$ 的原因；如果选择一个正数，相应的解将是一个关于 t 的指数增长的函数，而模型问题的解对任何正值 t 都是一致有界的. k 取任何值时上式都是微分方程的解. 如果限制 k 取值 $k=m\pi$, m 取正整数，解将在 $x=0$ 和 $x=1$ 处均取

零值. 因此, 所有这种解的线性组合均满足原微分方程和两个边界条件. 这种线性组合可以写为

$$u(x,t)=\sum_{m=1}^{\infty}a_m\mathrm{e}^{-(m\pi)^2t}\sin m\pi x. \tag{2.11}$$

现在必须选择适当的系数 a_m, 使这个线性组合满足给定的初始条件. 取 $t=0$ 得到

$$\sum_{m=1}^{\infty}a_m\sin m\pi x=u^0(x). \tag{2.12}$$

不难看出 a_m 恰是给定函数 $u^0(x)$ 的傅里叶正弦级数展开的系数, 因此得到

$$a_m=2\int_0^1 u^0(x)\sin m\pi x\mathrm{d}x. \tag{2.13}$$

这个最终结果可以认为是真解的解析表达式, 但从两个方面来说, 它更像一种数值逼近. 如果要得到 $u(x,t)$ 在特定 x 和 t 值处的函数值, 首先需要确定傅里叶系数 (Fourier coefficient) a_m; 而只有少数非常简单的 $u^0(x)$ 可以得到准确值, 更一般的需要进行某种形式的数值积分. 其次, 只能对无穷级数中的有限项求和. 但是, 这个方法对模型问题仍然是非常有效的; 即便在 t 值很小的情况下, 由于这个级数收敛速度非常快, 求出级数中的前几项就足够精确了. 这种形式的分离变量法的真正局限性在于它很难推广到稍微复杂些的微分方程.

2.4 模型问题的显式格式

用分别平行于 x 轴和 t 轴的两族直线在闭区域 $\bar{R}\times[0,t_F]$ 上划分网格, 使我们可以用有限差分方法逼近模型方程 (2.7). 为简单起见, 假定这些直线是等间距的, 并且从现在开始将 $\bar{R}$ 取为区间 $[0,1]$. 虽然在实际情况中只能在有限时间范围 $[0,t_F]$ 内计算, 但是 t_F 可取任意大的数值.

记 Δx 和 Δt 分别为两族直线间的间距. 相交点

$$(x_j=j\Delta x,\ t_n=n\Delta t),\ j=0,1,\ldots,J,\ n=0,1,\ldots, \tag{2.14}$$

称为**网格点** (grid point 或 mesh point), 其中

$$\Delta x=1/J. \tag{2.15}$$

我们在网格点上逼近方程的解; 记逼近值为

$$U_j^n\approx u(x_j,t_n). \tag{2.16}$$

首先用有限差分逼近 (2.7) 中的导数, 然后从 $n=0$ 开始逐步求解得到的差分方程组.

经常会用到的符号如 U_j^n, 它们一般不会和其他相似的表达式混淆, 例如 λ 的 n 次幂

λ^n. 但是如果出现任何可能的歧义，应记幂的形式为 $(\lambda_j)^n$.

对于模型问题，最简单的差分格式是在网格点 (x_j, t_n) 上向前差分逼近时间方向的导数，对任意有连续时间导数的函数 v 有

$$\frac{v(x_j, t_{n+1}) - v(x_j, t_n)}{\Delta t} \approx \frac{\partial v}{\partial t}(x_j, t_n); \tag{2.17}$$

而对二阶空间导数采用二阶中心差分

$$\frac{v(x_{j+1}, t_n) - 2v(x_j, t_n) + v(x_{j-1}, t_n)}{(\Delta x)^2} \approx \frac{\partial^2 v}{\partial x^2}(x_j, t_n). \tag{2.18}$$

令 (2.17) 和 (2.18) 的左端相等，则得到差分逼近所满足的方程

$$U_j^{n+1} = U_j^n + \mu\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right), \tag{2.19}$$

其中

$$\mu := \frac{\Delta t}{(\Delta x)^2}. \tag{2.20}$$

方程 (2.19) 中涉及的网格点的模式如图 2-1 所示． 显然，时间层 t_{n+1} 上的每个逼近值可

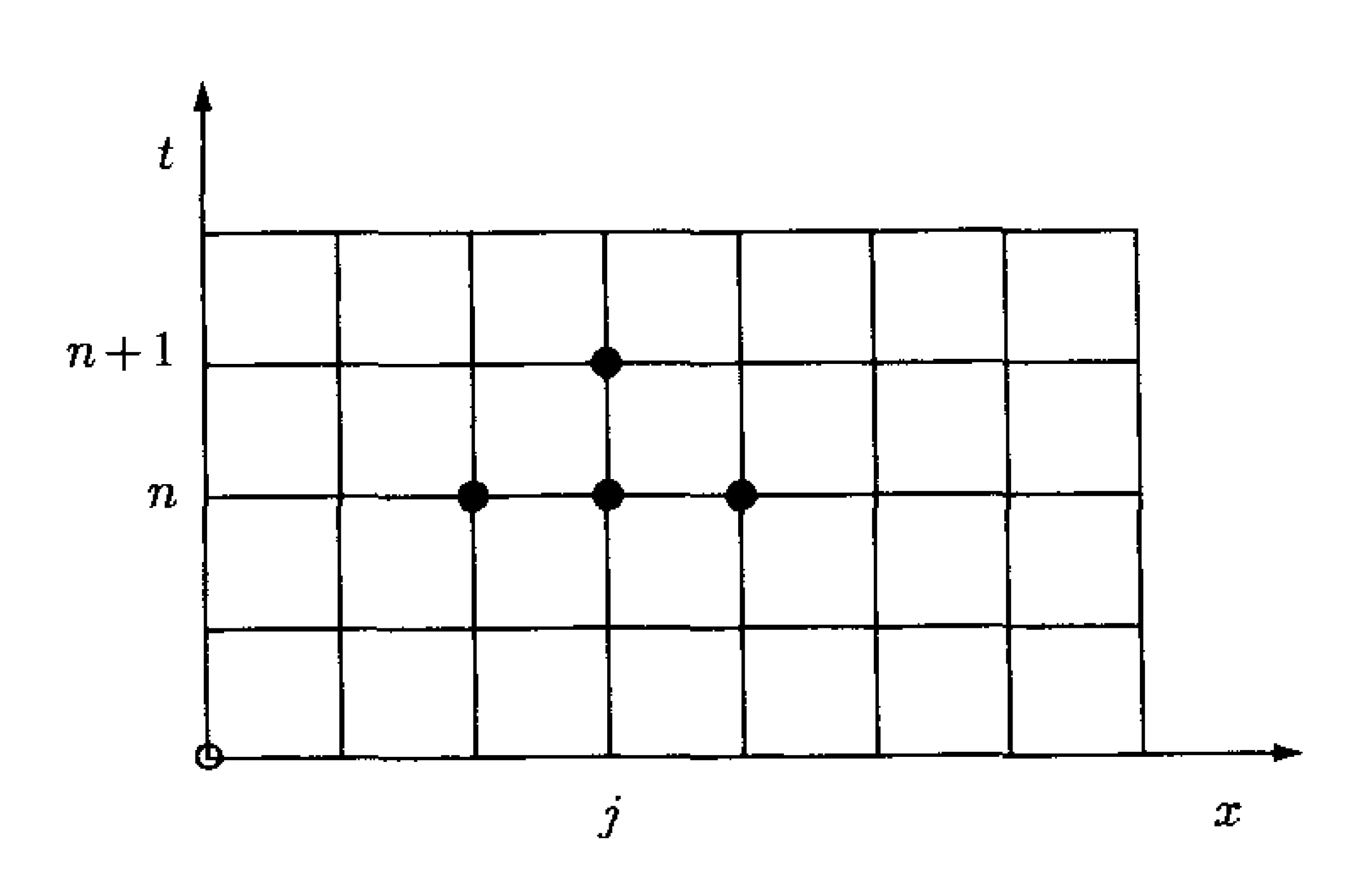


图 2-1 一种显式格式

以互相独立地由 t_n 层上的值求出；因此这个格式被称为*显式差分格式* (explicit difference scheme). 从初始值和边界值

$$U_j^0 = u^0(x_j),\ j = 1, 2, \ldots, J-1, \tag{2.21}$$

$$U_0^n = U_J^n = 0,\ n = 0, 1, 2, \ldots, \tag{2.22}$$

出发，按照 n 的次序，我们可以计算所有内部网格点上的值. 暂且假定初始值和边界值在两个角点处是相容的，亦即

$$u^0(0) = u^0(1) = 0. \tag{2.23}$$

因此解在区域的角点处不会出现不连续性.

可是，如果利用 (2.19), (2.21) 和 (2.22) 计算，很快就会发现数值结果严重依赖于 μ 值的选取，即它与时间和空间步长的相对大小有关. 图 2-2 中显示的是以“帽形函数”

$$u^0(x)=\begin{cases}2x, & \text{当}\quad 0\leqslant x\leqslant \frac{1}{2},\\ 2-2x, & \text{当}\quad \frac{1}{2}\leqslant x\leqslant 1.\end{cases} \tag{2.24}$$

为初始值得到的数值结果. 图中显示的是两组结果，均取 $J=20$, $\Delta x=0.05$. 第一组取 $\Delta t=0.0012$, 第二组取 $\Delta t=0.0013$. 前者的结果明显比较准确；而后者则出现振荡，并且随着时间 t 的增长越来越剧烈. 这是一个典型的 **稳定性** 或 **不稳定性** (stability 或 instability) 依赖网格比 μ 的例子. 两组数值解的不同是非常明显的. 它们虽然使用了几乎相等的时间步，但是 μ 的不同选取足以引起数值结果的形式完全不同.

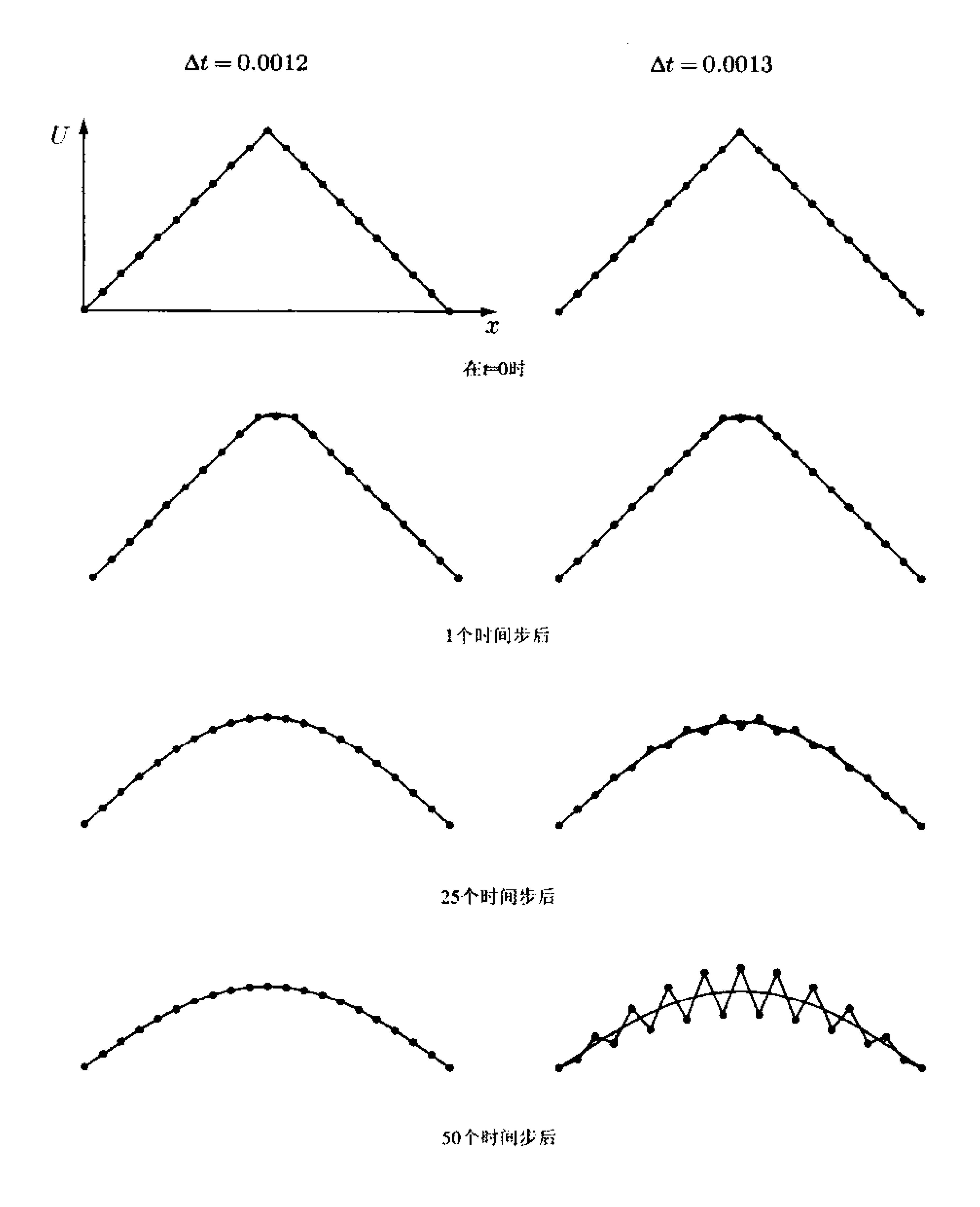


图 2-2 用显式方法和数据 (2.24) 得到的结果； $J=20, \Delta x=0.05$. 准确解是图中的实曲线

下面我们来分析这些现象，并且以更严格的方式得到误差界. 首先介绍一些定义和概念.

2.5　差分格式和截断误差

以相同的方式对变量 t 和 x 定义有限差分；这里有三种类型：

向前差分 (forward difference)

$$\Delta_{+t}v(x,t) := v(x,t+\Delta t) - v(x,t), \tag{2.25a}$$

$$\Delta_{+x}v(x,t) := v(x+\Delta x,t) - v(x,t); \tag{2.25b}$$

向后差分 (backward difference)

$$\Delta_{-t}v(x,t) := v(x,t) - v(x,t-\Delta t), \tag{2.26a}$$

$$\Delta_{-x}v(x,t) := v(x,t) - v(x-\Delta x,t); \tag{2.26b}$$

中心差分 (central difference)

$$\delta_t v(x,t) := v(x,t+\frac{1}{2}\Delta t) - v(x,t-\frac{1}{2}\Delta t), \tag{2.27a}$$

$$\delta_x v(x,t) := v(x+\frac{1}{2}\Delta x,t) - v(x-\frac{1}{2}\Delta x,t). \tag{2.27b}$$

连续应用两次中心差分算子，得到常用的二阶中心差分

$$\delta_x^2 v(x,t) := v(x+\Delta x,t) - 2v(x,t) + v(x-\Delta x,t). \tag{2.28}$$

对于一阶差分，用两个区间的中心差分公式

$$\begin{aligned}\Delta_{0x}v(x,t) \quad &:= \quad \frac{1}{2}\left(\Delta_{+x} + \Delta_{-x}\right)v(x,t)\\ &= \quad \frac{1}{2}\left[v(x+\Delta x,t) - v(x-\Delta x,t)\right].\end{aligned}$$

是比较方便的.

若 u 为 (2.7) 的解，对关于 t 的向前差分做泰勒级数展开

$$\begin{aligned}\Delta_{+t}u(x,t) \quad &= \quad u(x,t+\Delta t) - u(x,t)\\ &= \quad u_t\Delta t + \frac{1}{2}u_{tt}(\Delta t)^2 + \frac{1}{6}u_{ttt}(\Delta t)^3 + \cdots.\end{aligned} \tag{2.29}$$

把 $\Delta_{+x}u$ 和 $\Delta_{-x}u$ 关于变量 x 的泰勒级数展式相加， Δx 的奇次幂相互抵消，得到

$$\delta_x^2 u(x,t) = u_{xx}(\Delta x)^2 + \frac{1}{12}u_{xxxx}(\Delta x)^4 + \cdots. \tag{2.30}$$

现在可以定义格式 (2.19) 的*截断误差* (truncation error). 为了让差分方程中每一项是微分方程中相应导数的逼近，首先给差分方程乘以适当的因子. 若不用 (2.19) 而使用形式

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{(\Delta x)^2}, \tag{2.31}$$

则不必做这一步．截断误差就是用微分方程的准确解 $u(x_j,t_n)$ 代替近似值 U_j^n 后，差分方程两边的差．其实，在除了边界点以外的任意点都可以定义

截断误差 $T(x,t)$

$$T(x,t) := \frac{\Delta_{+t}u(x,t)}{\Delta t} - \frac{\delta_x^2 u(x,t)}{(\Delta x)^2}, \tag{2.32}$$

注意 u 满足微分方程，便得到

$$\begin{aligned} T(x,t) &= (u_t - u_{xx}) + \left(\frac{1}{2}u_{tt}\Delta t - \frac{1}{12}u_{xxxx}(\Delta x)^2\right) + \cdots \\ &= \frac{1}{2}u_{tt}\Delta t - \frac{1}{12}u_{xxxx}(\Delta x)^2 + \cdots, \end{aligned} \tag{2.33}$$

其中前面这几项 (leading terms) 被称为截断误差的 *主项* (principal part).

刚才用泰勒级数把截断误差表示为无穷级数．但是通过用余项把无穷泰勒级数截断为有穷项更为方便；例如

$$\begin{aligned} u(x,t+\Delta t) &= u(x,t) + u_t\Delta t + \frac{1}{2}u_{tt}(\Delta t)^2 + \frac{1}{6}u_{ttt}(\Delta t)^3 + \cdots \\ &= u(x,t) + u_t\Delta t + \frac{1}{2}u_{tt}(x,\eta)(\Delta t)^2, \end{aligned} \tag{2.34}$$

其中 η 位于 t 和 $t+\Delta t$ 之间．如果我们对 x 的展式作同样的处理，则截断误差变为

$$T(x,t) = \frac{1}{2}u_{tt}(x,\eta)\Delta t - \frac{1}{12}u_{xxxx}(\xi,t)(\Delta x)^2, \tag{2.35}$$

其中 $\xi \in (x-\Delta x, x+\Delta x)$. 由此得到

$$|T(x,t)| \leqslant \frac{1}{2}M_{tt}\Delta t + \frac{1}{12}M_{xxxx}(\Delta x)^2 \tag{2.36}$$

$$= \frac{1}{2}\Delta t\left[M_{tt} + \frac{1}{6\mu}M_{xxxx}\right], \tag{2.37}$$

其中 M_{tt} 和 M_{xxxx} 分别是 $|u_{tt}|$ 和 $|u_{xxxx}|$ 的上界． 我们当初假设 u 的初始值和边界值是相容的原因现在就清楚了，而且初始值足够光滑的假设也是有益的．因为这样我们可以假设在闭区域 $[0,1]\times[0,t_F]$ 上 M_{tt} 和 M_{xxxx} 一致有界． 否则就必须依赖扩散算子的光滑作用来确保对任意的 $\tau>0$ 在区域 $[0,1]\times[\tau,t_F]$ 上这些量是一致有界的．在许多看上去很直接的问题中都很容易出现这种困难．例如，边界条件要求 u 在边界 $x=0$ 和 $x=1$ 处取值为零，而在初始时刻 $t=0$ 时 u 取值 1. 这样解 $u(x,t)$ 在角点处必然不连续，并且在整个区域 $0<x<1$, $t>0$ 上它的所有导数都是无界的，因此我们在整个区域上求截断误差的界就没有意义了．后面我们将会看到如何用傅里叶分析处理这个问题．

对图 2-2 中的问题，可以看出，无论两个网格参数的关系如何，都有

$$T(x,t) \to 0, \quad \text{当 } \Delta x,\ \Delta t \to 0 \text{ 时} \quad \forall\ (x,t) \in (0,1)\times[\tau,t_F).$$

我们称这种格式与偏微分方程是*无条件相容的* (unconditionally consistent). 取固定比值 μ, 从 (2.37) 可以看出随 $\Delta t \to 0$, $|T|$ 与 $O(\Delta t)$ 渐近等价: 且除 μ 取特殊值之外, 这是使得这种渐近等价关系成立的 Δt 的最高阶次. 因此称这个格式有*一阶精度* (first order accuracy).

然而, 值得注意的是, 因为 u 处处满足 $u_t = u_{xx}$, 因此也满足 $u_{tt} = u_{xxxx}$, 故有

$$T(x,t) = \frac{1}{2}\left(1 - \frac{1}{6\mu}\right) u_{xxxx}\Delta t + O\left((\Delta t)^2\right).$$

即当 $\mu = \frac{1}{6}$ 时格式是*二阶精度* (second order accurate). 这当然是一种非常特殊的情况. 不但仅适用于特殊选择的 μ , 而且对更一般的变系数方程也是不成立的. 例如, 在求解方程 $u_t = b(x,t)u_{xx}$ 时, 它将要求在每一点选择不同的时间步长 Δt.

2.6 显式格式的收敛性

现在假设用相同的初值和相同的 $\mu = \Delta t/(\Delta x)^2$ 进行一系列计算. 但是相继加密两个方向的网格, 以使 $\Delta t \to 0$ 和 $\Delta x \to 0$. 在给定区域 $(0,1)\times(\tau, t_F)$ 中任意固定点 (x^*, t^*) 上:

$$\text{当} \quad x_j \to x^*, \ t_n \to t^* \quad \text{时}, \quad \text{有} \quad U_j^n \to u(x^*, t^*). \tag{2.38}$$

则称格式是*收敛的*(convergent). 我们将证明 $\mu \leqslant \frac{1}{2}$ 时前述问题的显式格式是收敛的.

我们仅需要考虑 (x^*, t^*) 是网格点的情况, 因为在网格足够细的情况下由 $u(x,t)$ 的连续性可得到其他点的收敛性. 在给定的网格上, 假定可以引入一个截断误差的上界 $\bar{T} = \bar{T}(\Delta x, \Delta t)$, 即在所有网格点上都有

$$|T_j^n| \leqslant \bar{T}, \tag{2.39}$$

其中符号 T_j^n 表示 $T(x_j, t_n)$. 用 e 表示逼近误差 $U - u$; 更准确地说

$$e_j^n := U_j^n - u(x_j, t_n). \tag{2.40}$$

U_j^n 严格满足方程 (2.19), 而 $u(x_j, t_n)$ 有余项 $T_j^n \Delta t$, 这可以从 T_j^n 的定义直接得到. 因此相减之后得到

$$e_j^{n+1} = e_j^n + \mu\delta_x^2 e_j^n - T_j^n \Delta t; \tag{2.41}$$

展开后有

$$e_j^{n+1} = (1-2\mu)e_j^n + \mu e_{j+1}^n + \mu e_{j-1}^n - T_j^n \Delta t. \tag{2.42}$$

证明的关键是验证当 $\mu \leqslant \frac{1}{2}$ 时方程右端 e^n 的三个系数都是正的, 并且其和为 1. 如果我们记在一个时间步上解的最大误差为

$$E^n := \max\{|e_j^n|, \ j = 0, 1, \ldots, J\}, \tag{2.43}$$

因为所有系数都是正值，所以在应用三角不等式运算时可以省略绝对值符号，于是得

$$
\begin{aligned}
|e_j^{n+1}| &\leqslant (1-2\mu)E^n + \mu E^n + \mu E^n + |T_j^n|\Delta t \\
&\leqslant E^n + \bar{T}\Delta t. \qquad (2.44)
\end{aligned}
$$

因为这个不等式对从 1 到 $J-1$ 的所有 j 都成立，所以有

$$E^{n+1} \leqslant E^n + \bar{T}\Delta t. \qquad (2.45)$$

暂且假定 (2.39) 中的界在有限区间 $[0, t_F]$ 上成立；并且 U_j^n 取给定的初始值，故有 $E^0 = 0$. 通过简单的归纳便得到 $E^n \leqslant n\bar{T}\Delta t$. 因此由 (2.37) 可得

$$E^n \leqslant \frac{1}{2}\Delta t\left[M_{tt} + \frac{1}{6\mu}M_{xxxx}\right]t_F \to 0, \text{ 当 } \Delta t \to 0 \text{ 时}. \qquad (2.46)$$

在我们的模型问题中，如果有必要还可以得到 $M_{tt} = M_{xxxx}$.

现在可以用更一般的术语叙述收敛性质．为了定义涉及两个网格尺度 Δt 和 Δx 的差分格式的收敛性，我们需要假定当它们都趋向于零时彼此之间的关系．因此引进加密路径的概念． *加密路径* (refinement path) 是一列由 Δt 和 Δx 组成的成对网格尺度， Δt 和 Δx 均趋向于零：

$$\text{加密路径} := \{((\Delta x)_i, (\Delta t)_i),\ i = 0,1,2,\ldots; (\Delta x)_i, (\Delta t)_i \to 0\}. \qquad (2.47)$$

可以按需要指定特殊的加密路径，例如要求 $(\Delta t)_i$ 和 $(\Delta x)_i$ 成比例，或者和 $(\Delta x)_i^2$ 成比例．在此我们定义

$$\mu_i = \frac{(\Delta t)_i}{(\Delta x)_i^2} \qquad (2.48)$$

并且仅要求 $\mu_i \leqslant \frac{1}{2}$. 图 2-3 给出了一些例子.

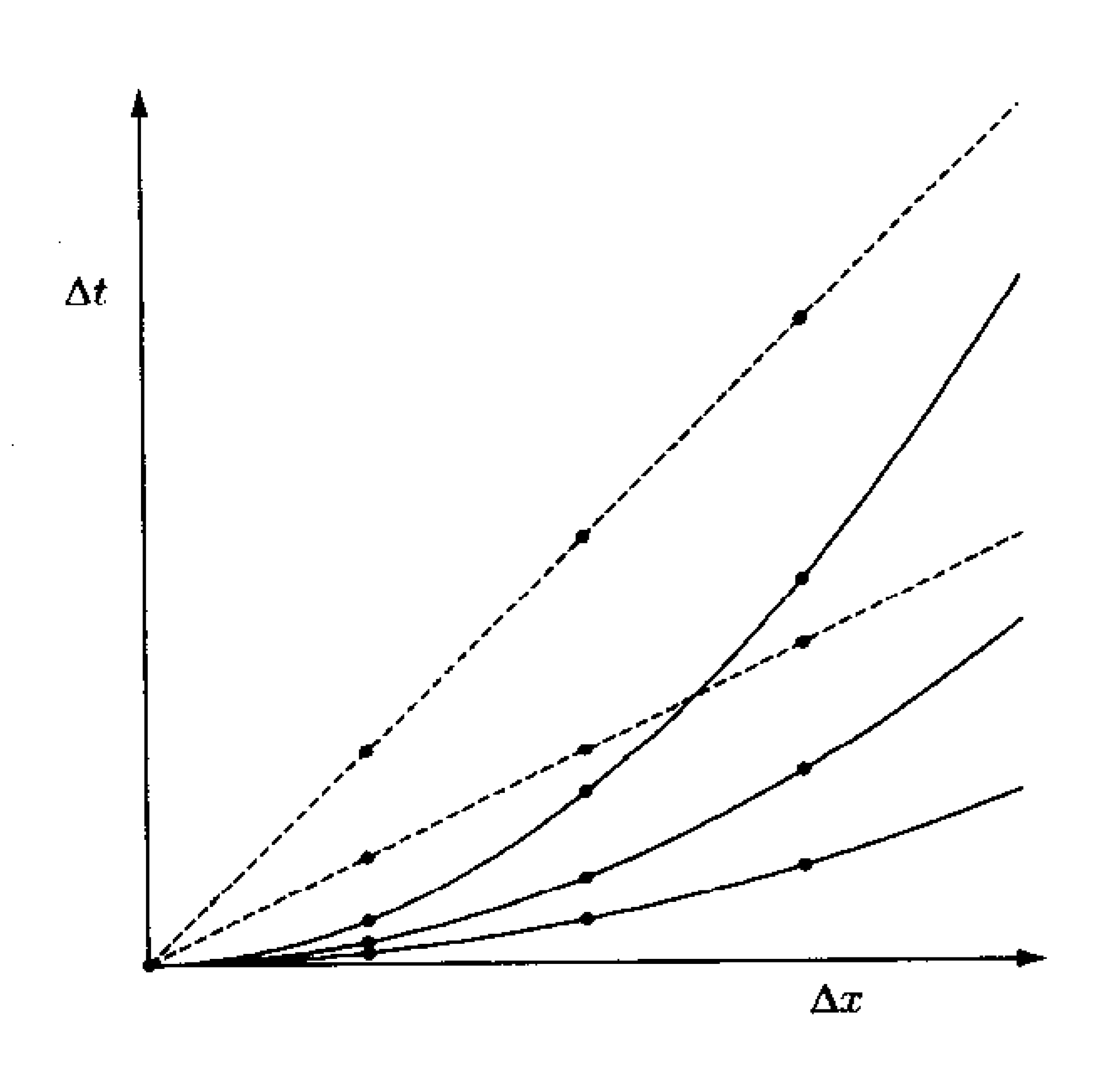


图 2-3　加密路径．实线表示 $\Delta t/(\Delta x)^2$ 是常数，虚线表示 $\Delta t/\Delta x$ 是常数

定理 2.1. 如果加密路径对所有充分大的 i 均满足 $\mu_i \leqslant \frac{1}{2}$, 正数 n_i, j_i 满足

$$n_i(\Delta t)_i \to t \ > \ 0, \quad j_i(\Delta x)_i \to x \ \in \ [0,1],$$

并且若在区域 $[0,1]\times[0,t_F]$ 上一致地有 $|u_{xxxx}| \leqslant M_{xxxx}$, 则由显式差分格式 (2.19) 得到的逼近 $U_{j_i}^{n_i}$, $i=0,1,2,\ldots$ 在该区域上一致收敛到微分方程的解 $u(x,t)$.

这个收敛定理是对数值格式的最低要求；它保证在足够细密的网格下得到的解可以达到足够高的精度. 当然，这有一点不切实际，随着网格的加密需要越来越多的计算量，同时计算中的舍入误差就变得非常大，最终它将完全淹没截断误差.

在区域 $[0,1]\times[0,1]$ 考虑热传导方程，初边值为

$$u(x,0) = x(1-x), \tag{2.49a}$$

$$u(0,\,t) = u(1,\,t) = 0. \tag{2.49b}$$

它比 (2.24) 所给的数据具有更好的光滑性. 由显式方法得到的误差如图 2-4 所示. 图中

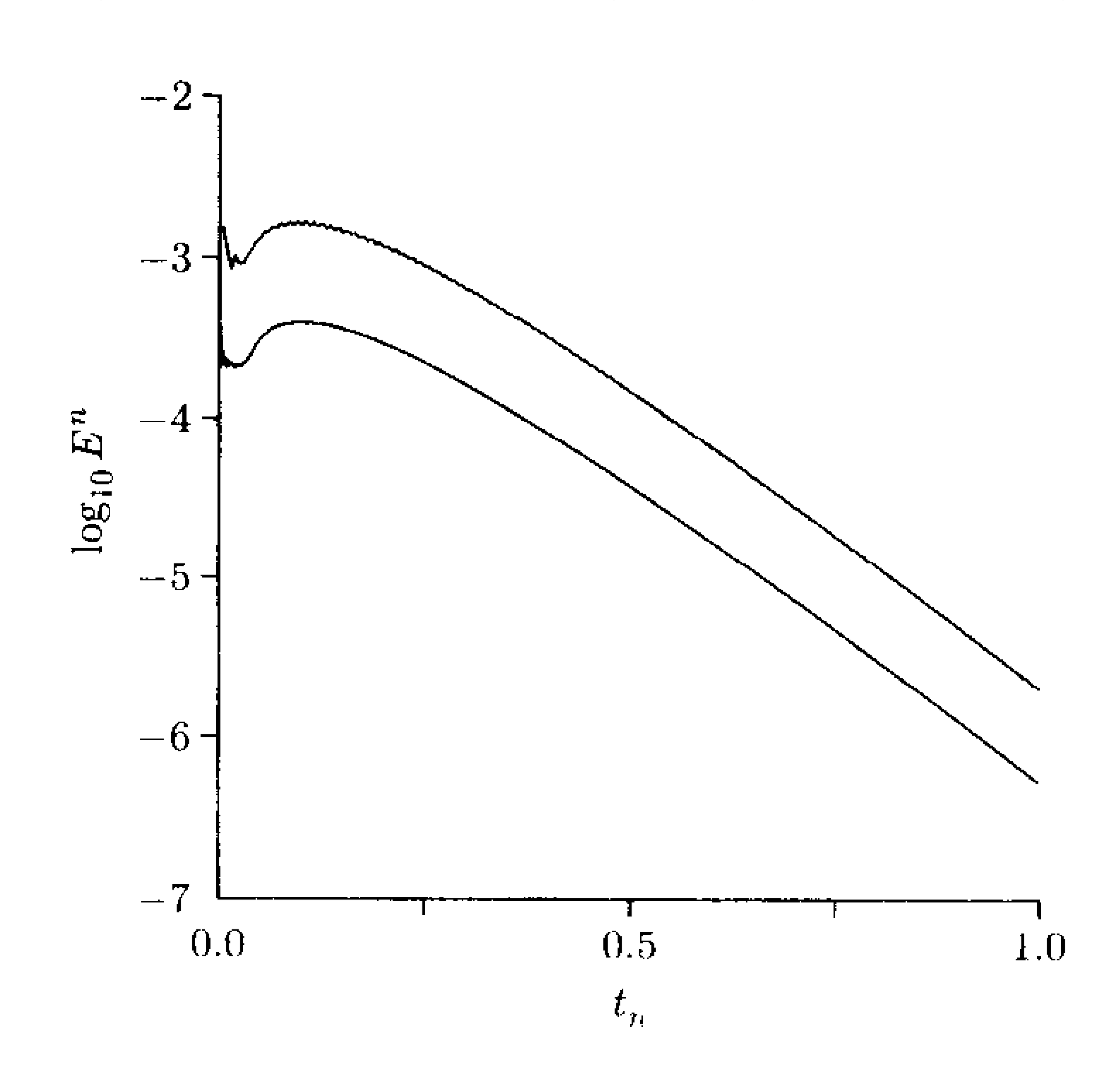


图 2-4　用显式方法求解热传导方程的误差衰减, 初值条件为 $u(x,0)=x(1-x)$. 上方的曲线取 $\Delta x = 0.1$, $\mu = 0.5$, 下方的曲线取 $\Delta x = 0.05$, $\mu = 0.5$

$\log_{10} E^n$ 为纵坐标，而 t_n 为横坐标， E^n 由 (2.43) 定义. 图中的两条曲线一条取 $J=10$, $\Delta x = 0.1$; 另一条取 $J=20$, $\Delta x = 0.05$. 两者均有 $\mu = \frac{1}{2}$, 也就是符合稳定性的最大值. 从两条曲线可以明显地看出误差随着网格尺度的减小的变化：它们的形状很相似，在每个 t_n 点，两个 E^n 的比值都接近于 4, 即 $\Delta t = \frac{1}{2}(\Delta x)^2$ 值的比率. 同时也注意到经过初期的摆动之后，误差随 t 的增长趋于零；而 (2.45) 中得到的误差界随着 t 的增长是增大的，所以这个界过于保守. 误差结果中初期所出现的摆动是由前面已经提及到的区域角点处的不光滑性造成的. 这一点我们将在下一节和 2.10 节更详细地讨论.

2.7 误差的傅里叶分析

基于一族特殊的傅里叶波型 (Fourier mode) 是方程精确解的认识，我们把偏微分方程的精确解表示为傅里叶级数的形式. 容易证明，类似傅里叶波型也是差分方程的精确解. 假如把

$$U_j^n = (\lambda)^n e^{i k(j\Delta x)} \tag{2.50}$$

代入差分方程 (2.19), 有 $U_j^{n+1} = \lambda U_j^n$, 其他项作类似处理. 然后除以 U_j^n, 这时易见，若

$$\begin{aligned} \lambda \equiv \lambda(k) &= 1 + \mu\left(e^{i k\Delta x} - 2 + e^{-i k\Delta x}\right) \\ &= 1 - 2\mu(1 - \cos k\Delta x) \\ &= 1 - 4\mu \sin^2 \frac{1}{2}k\Delta x, \end{aligned} \tag{2.51}$$

则对任意 n 和 j 这个傅里叶波型都是一个解； $\lambda(k)$ 称为此波型的增长因子(amplification factor). 与式 (2.11) 相同取 $k = m\pi$, 则数值逼近解可以用如下形式表示

$$U_j^n = \sum_{-\infty}^{\infty} A_m e^{i m\pi(j\Delta x)} \left[\lambda(m\pi)\right]^n. \tag{2.52}$$

由于 $\lambda(k)$ 和 $\exp(-k^2\Delta t)$ 两者的级数展开间存在很好的匹配关系：

$$\begin{aligned} \exp(-k^2\Delta t) &= 1 - k^2\Delta t + \frac{1}{2}k^4(\Delta t)^2 - \cdots, \\ \lambda(k) &= 1 - 2\mu\left[\frac{1}{2}(k\Delta x)^2 - \frac{1}{24}(k\Delta x)^4 + \cdots\right] \\ &= 1 - k^2\Delta t + \frac{1}{12}k^4\Delta t(\Delta x)^2 - \cdots. \end{aligned} \tag{2.53}$$

因此表达式中的低频项是微分方程 (2.11) 精确解的很好近似. 的确这些表达式提供了另一种研究格式截断误差的方法. 容易看出格式至少有一阶精度，并且当 $(\Delta x)^2 = 6\Delta t$ 时有二阶精度. 事实上，不难证明存在仅依赖于 μ 的常数 $C(\mu)$ 使得

$$|\lambda(k) - e^{-k^2\Delta t}| \leqslant C(\mu)k^4(\Delta t)^2, \qquad \forall k,\ \Delta t > 0. \tag{2.54}$$

定理 2.1 在条件 $\mu \leqslant \frac{1}{2}$ 下证明了收敛性并给出了误差界；但对不满足这个条件的情形没作任何说明. 对后者，傅里叶波型分析方法解释了其高频部分的变化. 因为含有指数项 $\exp(-k^2t)$, 所以精确解中 k 值很大的波型迅速衰减. 但在数值解中，如果 $\mu > \frac{1}{2}$ 且 k 值很大，衰减因子 $|\lambda(k)|$ 会大于 1; 特别地，当 $k\Delta x = \pi$ 时，将出现这种情况，因为这时 $\lambda(k) = 1 - 4\mu$. 这些傅里叶波型会随 n 的增大无界增长. 理论上可以选择合适的初值让这样的傅里叶波型不在解中出现，但是这只不过是理论上的特例，而在实际中舍入误差会引入振幅很小的各种波型项，其中的一部分将无界增长. 对目前的模型问题，如果存在

不依赖于 k 的常数 K 满足

$$|[\lambda(k)]^n| \leqslant K, \quad 对\ n\Delta t \leqslant t_F, \ \forall k. \tag{2.55}$$

则称方法是**稳定的** (stable). 本质上，稳定性关系到差分方程两解之间的差值在有限时间范围内关于网格尺寸一致有界的增长；在后面的章节中会给出一个一般的稳定性定义. 显然，稳定性要求 von Neumann 条件，即对所有的 k

$$|\lambda(k)| \leqslant 1 + K'\Delta t. \tag{2.56}$$

我们将看到相容的差分格式去逼近单个微分方程式时，如此定义的稳定性条件与收敛性是等价的. 对目前的模型问题，该方法在 $\mu > \frac{1}{2}$ 时是不稳定的， $\mu \leqslant \frac{1}{2}$ 时是稳定的.

由于容易与精确解比较，我们把 U_j^n 亦表示为一个无穷傅里叶级数 (2.52). 尽管如此，在离散网格上只有有限个不同的波型； 如果 $(k_1 - k_2)\Delta x$ 是 2π 的倍数，对应波数(wave number) k_1 和 k_2 的波型是不可区分的. 因此把 U_j^n 展成对应于

$$k = m\pi, \ m = -(J-1), -(J-2), \ldots, -1, 0, 1, \ldots, J \tag{2.57}$$

的不同波型的线性组合会更方便. $k = J\pi$ 或 $k\Delta x = \pi$ 是网格上可以表示的频率最高的波型；这个波型在网格点上交错地取值 ± 1. 从 (2.51) 可以看出，其增长因子是 $\lambda(J\pi) = 1 - 4\mu$，正像在许多差分格式中表现的一样，它是这个差分格式最不稳定的波型. 因为 $\mu > \frac{1}{2}$ 时它是增长最快的波型，所以如图 2-2 中所示，它最终在数值解中占据了主导地位.

我们还可以用此傅里叶分析方法将收敛性定理推广到初始值 $u^0(x)$ 在 $[0,1]$ 上连续但不光滑的情况，特别是处理角点的情况. 我们不再需要假设解有足够的有界导数，即 u_{xxxx} 和 u_{tt} 在考虑的区域上一致有界，而仅仅假设初值 $u^0(x)$ 的傅里叶级数展式是绝对收敛的. 我们假设 μ 固定且 $\mu \leqslant \frac{1}{2}$. 如前，考虑误差

$$\begin{aligned} e_j^n &= U_j^n - u(x_j, t_n) \\ &= \sum_{-\infty}^{\infty} A_m \mathrm{e}^{\mathrm{i}\, m\pi j\Delta x} \left\{ [\lambda(m\pi)]^n - \mathrm{e}^{-m^2\pi^2 n\Delta t} \right\}, \end{aligned} \tag{2.58}$$

其中使用了 $u(x,t)$ 的全部傅里叶级数而不是像特例 (2.11) 中那样仅用正弦级数；这样使我们可以处理比简单边界条件 (2.8) 更一般的条件. 现在把这个无穷和分为两部分. 对一任意正数 ϵ, 找到 m_0 满足

$$\sum_{|m|>m_0} |A_m| \leqslant \frac{1}{4}\epsilon. \tag{2.59}$$

因为级数是绝对收敛的，所以该不等式是可能成立的. 如果 $|\lambda_1| \leqslant 1$ 且 $|\lambda_2| \leqslant 1$, 则

$$|(\lambda_1)^n - (\lambda_2)^n| \leqslant n\,|\lambda_1 - \lambda_2|; \tag{2.60}$$

故由 (2.54) 可得

$$
\begin{aligned}
|e_j^n| &\leqslant \frac{1}{2}\epsilon + \sum_{|m|\leqslant m_0} |A_m| \left|[\lambda(m\pi)]^n - \mathrm{e}^{-m^2\pi^2 n\Delta t}\right| \\
&\leqslant \frac{1}{2}\epsilon + \sum_{|m|\leqslant m_0} |A_m| nC(\mu)\left(m^2\pi^2\Delta t\right)^2 .
\end{aligned}
\tag{2.61}
$$

因此可以推出

$$
|e_j^n| \leqslant \frac{1}{2}\epsilon + t_F C(\mu)\pi^4 \left[\sum_{|m|\leqslant m_0} |A_m| m^4\right] \Delta t. \tag{2.62}
$$

只要取 Δt 充分小，就可以使 $|e_j^n| \leqslant \epsilon$ 在 $[0,1]\times[0,t_F]$ 的所有 (x_j,t_n) 点上成立．注意正如前文分析 u_{xxxx} 的界一样，此时求和项 $A_m m^4$ 起到类似的作用．但是为了更准确的应用该格式的稳定性，我们不要求这个和式是收敛的．

2.8 隐式方法

稳定性条件 $\Delta t \leqslant \frac{1}{2}(\Delta x)^2$ 是非常严格的限制，并且意味着如果要在一个合理的时间长度上求解问题则需要计算非常多的时间步数．此外如果减小 Δx 来提高求解精度，则因为我们必须同时减小 Δt, 计算量的增长将非常迅速．如果我们用向后时间差分给出一个差分格式，则可以避免这种限制，但是其代价是计算过程稍微复杂些．

如果用向后时间差分代替向前时间差分，空间差分保持不变，得到的格式就是

$$
\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{(\Delta x)^2} \tag{2.63}
$$

而不是 (2.31)．使用 2.5 节中引入的差分符号可将上式表示为

$$
\Delta_{-t} U_j^{n+1} = \mu \delta_x^2 U_j^{n+1},
$$

其中 $\mu = \Delta t/(\Delta x)^2$, 且使用如图 2-5 所示的模板 (stencil).

这是一个*隐式格式*(implicit scheme) 的例子，它不像前面叙述的显式格式那样容易使用．格式 (2.63) 涉及 $n+1$ 时间层上 U 的三个未知值；因为方程包含同样未知的两个相邻的值 U_{j+1}^{n+1} 和 U_{j-1}^{n+1}, 我们不能立即由此计算出 U_j^{n+1}. 现在方程必须写为如下形式

$$
-\mu U_{j-1}^{n+1} + (1+2\mu)U_j^{n+1} - \mu U_{j+1}^{n+1} = U_j^n. \tag{2.64}
$$

让 j 取遍 $1, 2, \ldots, (J-1)$, 便得到一个关于 $J-1$ 个变量 U_j^{n+1} 的由 $J-1$ 个线性方程组成的方程组．我们现在必须通过求解这个方程组同时得到所有的未知值，而不能用简单的单个公式分别求解每个未知值．注意，在方程组中分别对应 $j=1$ 和 $j=J-1$ 的第一和

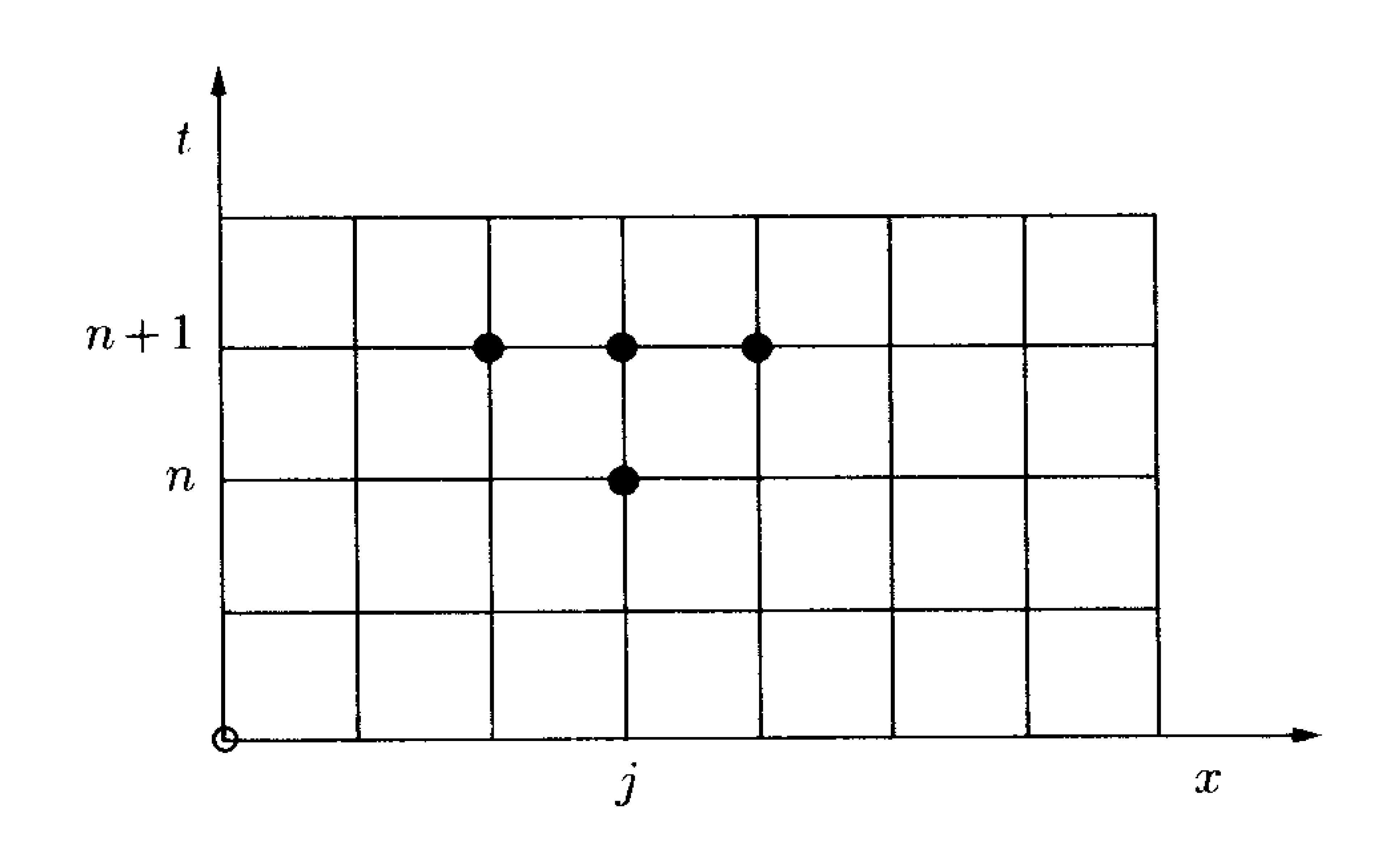


图 2-5 全隐式格式

最后一个方程中，我们必须代入由边界条件确定的已知值 U_0^{n+1} 和 U_J^{n+1}.

2.9 Thomas 算法

需要求解的方程组是三对角的：系统中第 j 个方程仅含有下标为 $j-1, j$ 和 $j+1$ 的未知值，所以方程组的系数矩阵只有对角元素及其左右两边相邻的元素非零．我们还会再遇到这类方程组，所以在此考虑具有以下形式的更具有一般性的方程组是有用的．

$$-a_jU_{j-1}+b_jU_j-c_jU_{j+1}=d_j, \quad j=1,2,\ldots,J-1, \tag{2.65}$$

且有

$$U_0=0, \quad U_J=0. \tag{2.66}$$

这里我们暂时忽略上标，把未知值记为 U_j. 系数 a_j, b_j 和 c_j, 和右端项 d_j 已知，并且假设它们满足条件

$$a_j>0,\ b_j>0,\ c_j>0, \tag{2.67}$$

$$b_j>a_j+c_j. \tag{2.68}$$

这些条件可以保证矩阵是对角占优的，即在每行中对角元素不小于其他元素的绝对值之和，虽然这里的条件比必要的条件更强，但是容易看出我们的差分方程组满足这些条件.

Thomas 算法是在每个方程中消去未知元素 U_{j-1}, 把方程组化为上三角形式．对每个方程依次进行这种消元运算．假设方程组 (2.65) 的前 k 个已经化为

$$U_j-e_jU_{j+1}=f_j, \quad j=1,2,\ldots,k. \tag{2.69}$$

故这些方程中最后一个为 $U_k-e_kU_{k+1}=f_k$, 且下一个仍为它的原始形式，即

$$-a_{k+1}U_k+b_{k+1}U_{k+1}-c_{k+1}U_{k+2}=d_{k+1}.$$

可以很容易地从这两个方程中消去 U_k, 得到一个关于 U_{k+1} 和 U_{k+2} 的新方程,

$$U_{k+1} - \frac{c_{k+1}}{b_{k+1} - a_{k+1}e_k}U_{k+2} = \frac{d_{k+1} + a_{k+1}f_k}{b_{k+1} - a_{k+1}e_k}.$$

与 (2.69) 比较，系数 e_j 和 f_j 可以利用递归关系

$$e_j = \frac{c_j}{b_j - a_je_{j-1}}, \quad f_j = \frac{d_j + a_jf_{j-1}}{b_j - a_je_{j-1}}, \quad j = 1, 2, \ldots, J-1; \tag{2.70}$$

计算；同时，令 $j = 0$, 把边界条件 $U_0 = 0$ 代入 (2.69), 得到初始值

$$e_0 = f_0 = 0. \tag{2.71}$$

利用递归关系求得这些系数后，U_j 很容易从 (2.69) 求出：已知 U_J, 由方程依次求出 U_{J-1}, $U_{J-2}, \ldots$, 最后求出 U_1.

用如 (2.69) 的递归关系相继求解 U_j 的过程一般来说数值是不稳定的，会导致误差递增. 尽管如此，如果对每个 j (2.69) 的系数均满足 $|e_j| < 1$, 则这种不稳定便不会发生，而条件 (2.67) 和 (2.68) 是保证其成立的充分条件，我们把它的证明留做习题 (见习题 2.4).

这个算法是非常有效的 (在串行计算机上); 每个网格点只需要 3(加) + 3(乘) + 2(除) 次运算就可以求解 (2.64), 与此相比显式格式 (2.19) 每个网格点需要 3(加) + 2(乘) 次运算. 因此每个时间步需要大约 (2.19) 两倍的计算量. 当然，隐式方法的重要性在于其时间步长可以非常大，因为正如将要看到的，稳定性对 Δt 不再有任何限制. 作为更一般方法的特例，下一节我们将会证明这个隐式格式的收敛性. 首先，可以用 2.7 节中的傅里叶方法验证它的稳定性.

如前，为差分方程构造一个傅里叶波型形式的解,

$$U_j^n = (\lambda)^n e^{i k(j\Delta x)}. \tag{2.72}$$

上式满足 (2.64), 若

$$\begin{aligned} \lambda - 1 &= \mu\lambda\left(e^{i k\Delta x} - 2 + e^{-i k\Delta x}\right) \\ &= -4\mu\lambda\sin^2\frac{1}{2}k\Delta x, \end{aligned} \tag{2.73}$$

即

$$\lambda = \frac{1}{1 + 4\mu\sin^2\frac{1}{2}k\Delta x}. \tag{2.74}$$

成立. 显然对任意正数 μ 都有 $0 < \lambda < 1$, 因此这个隐式方法是*无条件稳定的* (unconditionally stable). 在下一节我们会看到，其截断误差和显式格式的是同一数量级的，但这里不再需要对 μ 加任何约束条件以保证所有傅里叶波型都不会随 n 的增大而增长.

时间步仍受截断误差必须很小的限制，但是实际中发现在大多数问题中隐式格式的时间步 Δt 可以比显式格式的时间步大的多；虽然前者每一步需要大约后者两倍的工作

量，但是达到时间 t_F 的全部工作量却少许多.

2.10 加权平均和 θ-方法

现在我们已经考虑了两种有限差分方法，它们的不同在于一个用旧时间层 t_n 上的三点来逼近二阶空间导数，另一个则用的是新时间层 t_{n+1} 上的三点. 自然地它们可以推广到使用全部六点的逼近. 可以认为这是两个公式的加权平均. 因为左端的时间差分是一样的，我们得到一个六点格式 (见图 2-6)

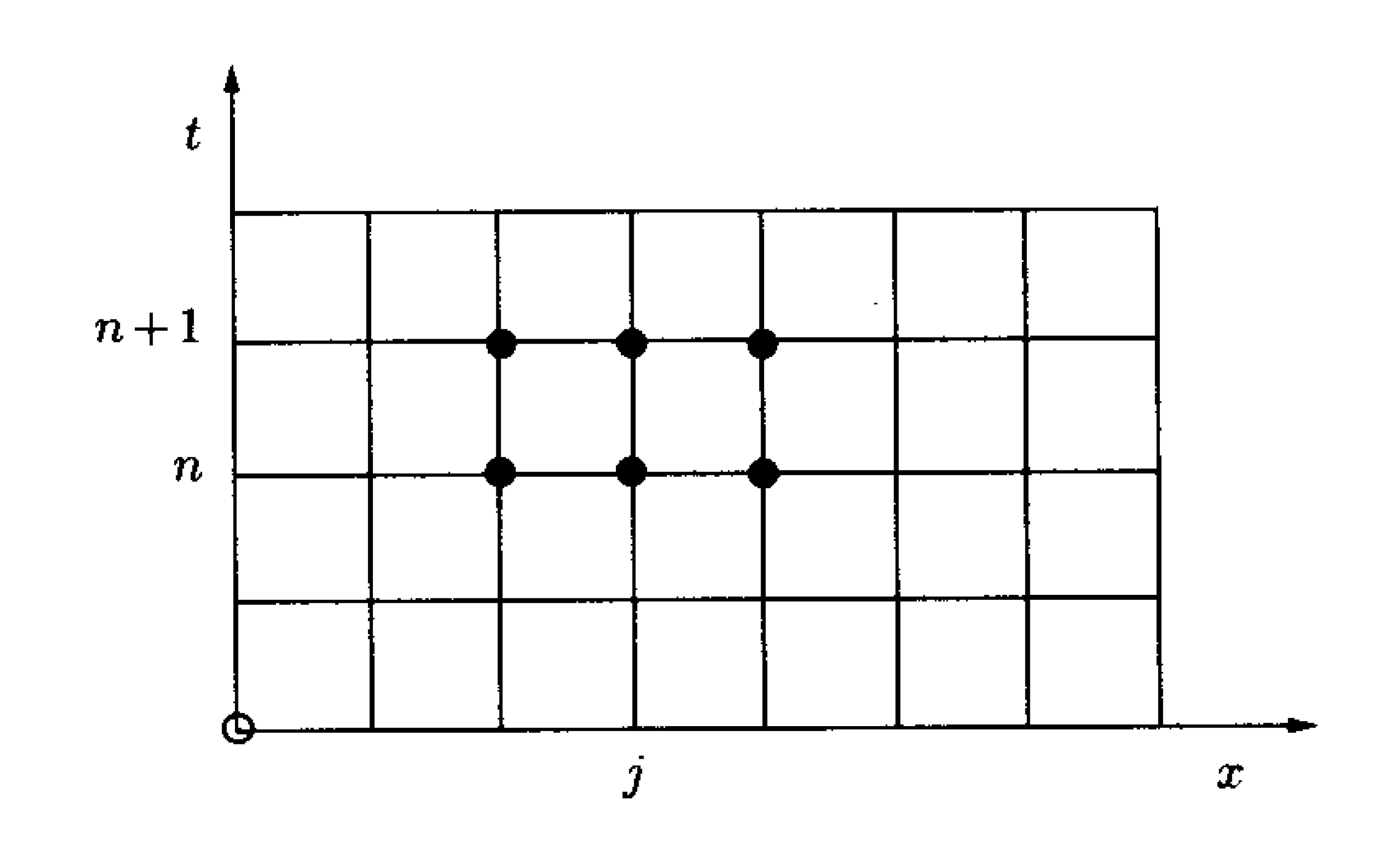


图 2-6 θ-方法

$$U_j^{n+1} - U_j^n = \mu\left[\theta\delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n\right], \quad j = 1, 2, \ldots, J-1. \tag{2.75}$$

假定我们使用取非负权的求平均方法，即 $0 \leqslant \theta \leqslant 1$; $\theta = 0$ 得到显式格式，$\theta = 1$ 得到全隐式格式. 对任意 $\theta \neq 0$, 为求解 $\{U_j^{n+1}\}$ 我们得到一个三对角方程组，即

$$-\theta\mu U_{j-1}^{n+1} + (1+2\theta\mu)U_j^{n+1} - \theta\mu U_{j+1}^{n+1} = \left[1 + (1-\theta)\mu\delta_x^2\right] U_j^n. \tag{2.76}$$

显然系数满足 (2.67) 和 (2.68), 所以方程组可以用前面求解全隐式格式所用的 Thomas 算法稳定求解.

用 2.7 节和以上介绍的傅里叶分析方法考虑这个单参数格式族的稳定性. 把波型 (2.72) 代入方程 (2.75), 得到

$$\begin{aligned}\lambda - 1 &= \mu\left[\theta\lambda + (1-\theta)\right]\left(\mathrm{e}^{\mathrm{i}k\Delta x} - 2 + \mathrm{e}^{-\mathrm{i}k\Delta x}\right)\\ &= \mu\left[\theta\lambda + (1-\theta)\right]\left(-4\sin^2\frac{1}{2}k\Delta x\right),\end{aligned}$$

即

$$\lambda = \frac{1 - 4(1-\theta)\mu\sin^2\frac{1}{2}k\Delta x}{1 + 4\theta\mu\sin^2\frac{1}{2}k\Delta x}. \tag{2.77}$$

因为 $\mu > 0$, 且假设 $0 \leqslant \theta \leqslant 1$, 显然 $\lambda > 1$ 恒不成立，因此只有当 $\lambda < -1$ 时才可能产生不稳定，也就是当

$$1 - 4(1-\theta)\mu \sin^2 \frac{1}{2}k\Delta x < -\left[1 + 4\theta\mu \sin^2 \frac{1}{2}k\Delta x\right],$$

亦即当

$$4\mu(1-2\theta)\sin^2 \frac{1}{2}k\Delta x > 2.$$

上式左边取最大值时是最可能产生不稳定的波型：如前，这是振荡最快的波型，因为此时 $k\Delta x = \pi$. 如果

$$\mu(1-2\theta) > \frac{1}{2}. \tag{2.78}$$

则这是不稳定的波型. 它包括了前面显式格式的情形，即 $\theta = 0$; 并且亦可以看出 $\theta = 1$ 的全隐式格式对任何 μ 值是不可能不稳定的. 确实当 $\theta \geqslant \frac{1}{2}$ 时， μ 取任何值格式都是稳定的. 如果条件 (2.78) 满足，则对一固定时刻当 $\Delta t \to 0$ 亦即 $n \to \infty$ 时存在无界增长的波型；另一方面如条件 (2.78) 不满足，则对所有的波型 k 都有 $|\lambda(k)| \leqslant 1$, 所以没有波型是增长的，因而格式是稳定的. 这样我们可以把 (2.75) 稳定性的充分必要条件归纳如下

$$\left.\begin{array}{ll} \text{当 } 0 \leqslant \theta < \frac{1}{2} \text{ 时,} & \text{当且仅当 } \mu \leqslant \frac{1}{2}(1-2\theta)^{-1} \text{ 时稳定,} \\ \text{当 } \frac{1}{2} \leqslant \theta \leqslant 1 \text{ 时,} & \text{对所有的 } \mu \text{ 都稳定.} \end{array}\right\} \tag{2.79}$$

这两种情况经常分别被称为有条件稳定性和无条件稳定性. 只要 θ 不为零则必须求解一个三对角方程组，因此，除非使用更精确的 $0 < \theta < \frac{1}{2}$ 格式，否则表面上看它们并没有什么优势，因为它们仅仅是条件稳定的. 因此下面我们应该分析 (2.75) 的截断误差.

计算这样一个六点格式的截断误差时，选择合适的点做泰勒级数展开是很重要的. 显然不论如何选择展开点，截断误差的主项是不变的，但是计算的难易和简化却受到本质影响. 对显式格式 (2.31) 自然而方便的选择是 (x_j, t_n) 点，基于同样的理由对纯隐式格式 (2.63) 自然的展开点是 (x_j, t_{n+1}). 然而，当 θ 取任何中间值时，我们应该选取六个网格点的中心，即 $(x_j, t_{n+1/2})$, 而且经常把截断误差记为 $T_j^{n+1/2}$. 在格式中把各项按照对称的方式分组也是有帮助的，这样可以最大限度的利用各泰勒展式间的相互抵消. 对 u 和 U 使用同样的上标 / 下标符号，则相应于 (2.75) 中的各项，有

$$u_j^{n+1} = \left[u + \frac{1}{2}\Delta t\, u_t + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2 u_{tt} + \frac{1}{6}\left(\frac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots\right]_j^{n+\frac{1}{2}},$$

$$u_j^{n} = \left[u - \frac{1}{2}\Delta t\, u_t + \frac{1}{2}\left(\frac{1}{2}\Delta t\right)^2 u_{tt} - \frac{1}{6}\left(\frac{1}{2}\Delta t\right)^3 u_{ttt} + \cdots\right]_j^{n+\frac{1}{2}}.$$

如果把这两个泰勒展式相减，则展式中所有的偶次项相互抵消，得到

$$\delta_t u_j^{n+1/2} = u_j^{n+1} - u_j^n = \left[\Delta t\, u_t + \frac{1}{24}(\Delta t)^3 u_{ttt} + \cdots\right]_j^{n+\frac{1}{2}}. \tag{2.80}$$

又由 (2.30) 有

$$\delta_x^2 u_j^{n+1} = \left[(\Delta x)^2 u_{xx} + \frac{1}{12}(\Delta x)^4 u_{xxxx} + \frac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right]_j^{n+1}. \tag{2.81}$$

现在我们把这个展式中的每一项在点 $(x_j, t_{n+1/2})$ 展开，表示为 Δt 幂级数. 为了简单地表示这些展式，我们忽略它们的上标和下标，且默认所得到的展式中每项都是在该展开点求值. 于是有

$$\begin{aligned}
\delta_x^2 u_j^{n+1} \quad = \quad & \left[(\Delta x)^2 u_{xx} + \frac{1}{12}(\Delta x)^4 u_{xxxx} + \frac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right] \\
& + \frac{1}{2}\Delta t\left[(\Delta x)^2 u_{xxt} + \frac{1}{12}(\Delta x)^4 u_{xxxxt} + \cdots\right] \\
& + \frac{1}{2}(\frac{1}{2}\Delta t)^2\left[(\Delta x)^2 u_{xxtt} + \cdots\right] + \cdots.
\end{aligned}$$

$\delta_x^2 u_j^n$ 也有一个类似的展式：把两展式做线性组合得到

$$\begin{aligned}
&\theta\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n = \\
&\left[(\Delta x)^2 u_{xx} + \frac{1}{12}(\Delta x)^4 u_{xxxx} + \frac{2}{6!}(\Delta x)^6 u_{xxxxxx} + \cdots\right] \\
&\quad + (\theta - \frac{1}{2})\Delta t\left[(\Delta x)^2 u_{xxt} + \frac{1}{12}(\Delta x)^4 u_{xxxxt} + \cdots\right] \\
&\quad + \frac{1}{8}(\Delta t)^2(\Delta x)^2[u_{xxtt}] + \cdots.
\end{aligned} \tag{2.82}$$

为了清楚地看出所有相关项的形式，这里比平常计算截断误差主项时保留了更多的项. 此外我们还没有使用 u 是微分方程解的这个事实，所以 (2.80) 和 (2.82) 对任何足够光滑的函数都是成立的. 如果用这些展式计算截断误差，可以得到

$$T_j^{n+1/2} \quad := \quad \frac{\delta_t u_j^{n+1/2}}{\Delta t} - \frac{\theta\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n}{(\Delta x)^2} \tag{2.83}$$

$$\begin{aligned}
= \quad & [u_t - u_{xx}] + \left[(\frac{1}{2} - \theta)\Delta t\, u_{xxt} - \frac{1}{12}(\Delta x)^2 u_{xxxx}\right] \\
& + \left[\frac{1}{24}(\Delta t)^2 u_{ttt} - \frac{1}{8}(\Delta t)^2 u_{xxtt}\right] \\
& + \left[\frac{1}{12}\left(\frac{1}{2} - \theta\right)\Delta t(\Delta x)^2 u_{xxxxt} - \frac{2}{6!}(\Delta x)^4 u_{xxxxxx}\right].
\end{aligned} \tag{2.84}$$

至此还没有进行任何消去运算，仅仅把可以消去的项分组排列.

式 (2.84) 中的首项总是可以消去，因此验证了对任意的 θ 和 μ 相容性成立的事实. 从第二项可以看出一般我们有一阶精度 (关于 Δt), 但对称平均 $\theta=\frac{1}{2}$ 是个特例：这个取值给出了著名且常用的 *Crank-Nicolson 格式* (Crank-Nicolson scheme). 这个格式以它的两个作者命名，在 1947 年的一篇论文[1]中他们成功的把这个格式应用到染织问题中. 由于即便利用了微分方程，(2.84) 中第三项也不能消去，而是得到

$$T_j^{n+\frac{1}{2}}=-\frac{1}{12}\left[(\Delta x)^2u_{xxxx}+(\Delta t)^2u_{ttt}\right]_j^{n+\frac{1}{2}}+\cdots\ (\text{当 }\theta=\frac{1}{2}\text{ 时}),\tag{2.85}$$

从中可以看出对 Δt 和 Δx ,Crank-Nicolson 格式总是二阶精度的：这意味着可以利用格式额外的稳定性以选用较大的时间步长，例如 $\Delta x=O(\Delta t)$, 因为截断误差是 $O\left((\Delta t)^2\right)$, 因此可以经济地得到更高的精度.

另一种选择时常被认为是 2.5 节中讨论内容的推广. 将 θ 与 Δx 和 Δt 相联系以彻底消去 (2.84) 中第二项，因此取

$$\theta=\frac{1}{2}-(\Delta x)^2/12\Delta t,\tag{2.86}$$

即

$$\mu=\frac{1}{6(1-2\theta)},\tag{2.87}$$

但是注意，这要求 $(\Delta x)^2\leqslant 6\Delta t$ 以保证 $\theta\geqslant 0$. 这时得到的 θ 值小于 $\frac{1}{2}$, 但容易看出它满足条件 (2.79), 因此是稳定的. 对显式格式的情况来说 θ 取零，则 μ 化简为 $\frac{1}{6}$. 彻底消去 (2.84) 中第二项后得到的截断误差为

$$\begin{aligned}T_j^{n+\frac{1}{2}}&=-\frac{1}{12}\left[(\Delta t)^2u_{ttt}+\frac{1}{20}(\Delta x)^4u_{xxxxxx}\right]_j^{n+\frac{1}{2}}+\cdots\\(\text{当 }\theta&=\frac{1}{2}-\frac{1}{12\mu}\text{ 时}),\end{aligned}\tag{2.88}$$

其阶为 $O\left((\Delta t)^2+(\Delta x)^4\right)$. 因此可以再次在保持精度和稳定性的前提下使用很大的时间步长：例如，取 $\Delta t=\Delta x=0.1$ 时 $\theta=\frac{1}{2}-\frac{1}{120}$, 所以得到格式与 Crank-Nicolson 格式非常接近.

还有许多差分格式可以用来处理热流方程，在 Richtmyer 和 Morton(1967) (pp.189-91) 的书中就罗列了十四种格式. 尽管如此，目前在实际中应用最广的还是这种两时间层、三空间点的 (2.75) 格式，尽管在处理不同的问题时 θ 的最优选择各不相同，甚至对同一给定问题也没有对哪种格式是最优的达成广泛共识. 下一节，我们将考虑这些更一般方

[1] Crank, J. and Nicolson, P. (1947) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. Philos. Soc.* **43**, 50-67.

法的收敛性分析：但是首先给出对问题 (2.49) 采用隐式格式的结果，它和图 2-4 中所示的由显式格式得到的结果很相似，其中显示出 Crank-Nicolson 方法特别精确．图 2-7 上相对于不同格式的曲线族中最大值误差 E 为纵坐标而网格数 J 为横坐标：为了剔除在 t 很小时的瞬时现象，我们取

$$E := \max\left\{|e_j^n|,\ (x_j, t_n)\ \in\ [0,1]\times[0.1,1]\right\}.$$

开始取 $J = 10$; 对每一个隐式格式当 $\mu = \Delta t/(\Delta x)^2$ 取固定值时画一条实曲线，而当 $\nu = \Delta t/\Delta x$ 取固定值时则画一条点划线；注意后一种情况中所需时间步数的增长要慢的多． μ 和 ν 值的选择是根据 $J = 10$ 时得到相同的 Δt 的原则；这要求 $\mu = 10\nu$. 对显式格式仅有 $\mu = \frac{1}{2}$ 时的一条曲线，即可能得到稳定结果的最大值.

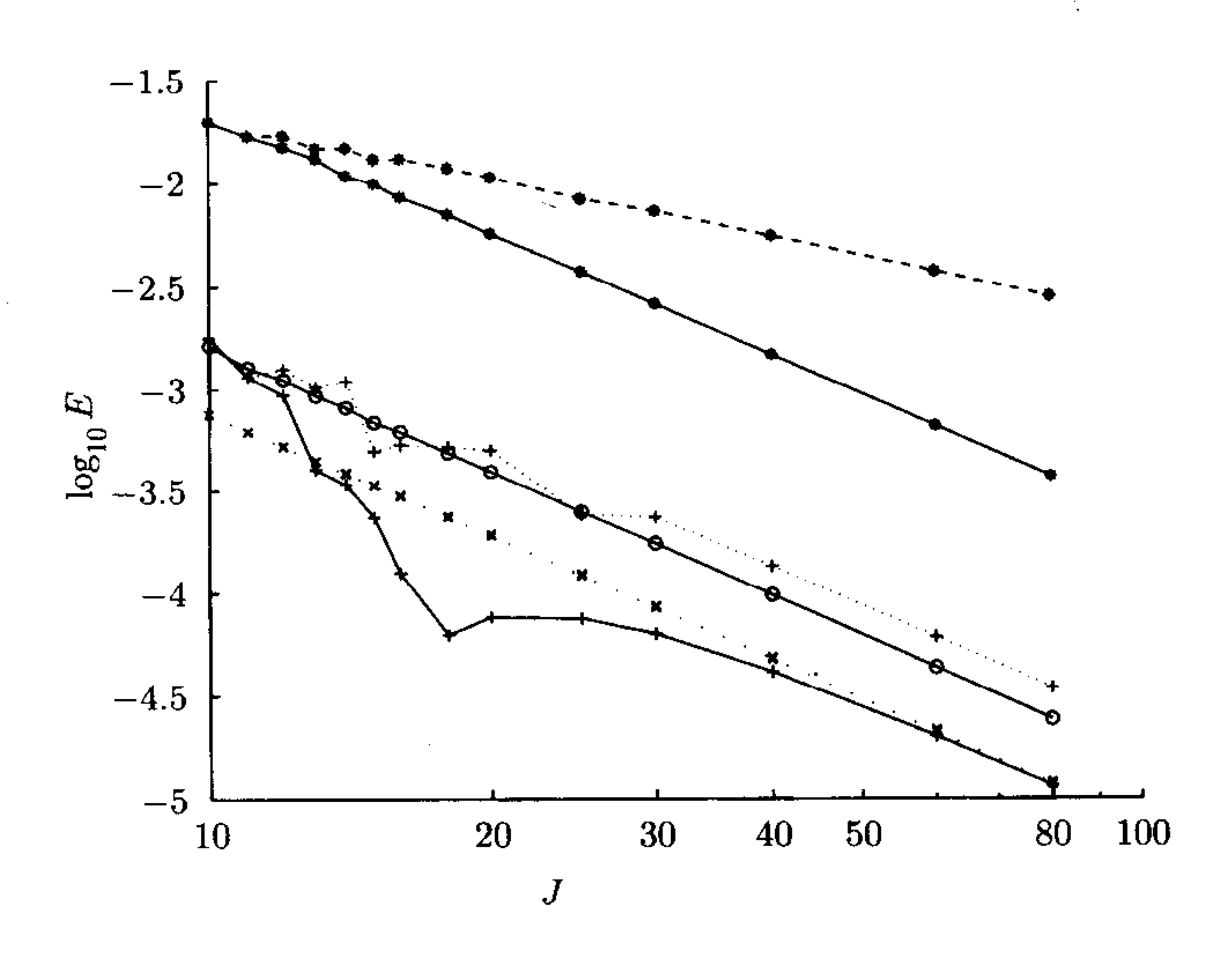


图 2-7　在区域 $[0,1]\times[0.1,1]$ 上，不同格式的最大值误差 E 相对于各网格数 J 的图

A :	$\theta = 0,\ \mu = \frac{1}{2}$	—○—○—○—○
B :	$\theta = \frac{1}{2},\ \mu = \frac{1}{2},$	—×—×—×—
	$\theta = \frac{1}{2},\ \nu = \frac{1}{20}$	···×···×···
C :	$\theta = \frac{1}{2},\ \mu = 5,$	—+—+—+—
	$\theta = \frac{1}{2},\ \nu = \frac{1}{2}$	···+···+···
D :	$\theta = 1,\ \mu = 5,$	—∗—∗—∗—
	$\theta = 1,\ \nu = \frac{1}{2}$	---∗---∗---∗---

图示 A 可以看出显式格式按所预计的 $O(\Delta t) = O(J^{-2})$ 变化；图示 B 中取 $\mu = \frac{1}{2}$ 的 Crank-Nicolson 格式也有同样现象．从 Crank-Nicolson 格式的截断误差表达式 (2.85) 可以

看出，当 μ 保持常值时，相对于第一项第二项是可以忽略的，所以两图示的差别仅在于截断误差之间的固定比例 $\frac{1}{2}$. 同样地，在下一节中将证明两种格式都满足最大值原理，因此它们的整体性态非常相似. 图示 B 取 $\nu=\frac{1}{20}$ 时，两项都是 $O\left((\Delta x)^2\right)$ 阶，但在数值上第二项比第一项小得多；这解释了图示 B 中两条曲线无法区分的原因.

图示 C 说明的也是 Crank-Nicolson 格式，但这里 $\mu=5$ 取值大得多. 当 t 很小时数值解出现振荡现象， C 中的两条曲线也因此更不规则，直到 J 取值大约 40 之后才会表现出所期望的性态. 当 J 值大于这个值以后，分别对应于 $\mu=\frac{1}{2}$ 和 $\mu=5$ 的两个曲线就非常接近，这说明在截断误差 (2.85) 中的主项与 μ 是无关的. 尽管如此，当 $\nu=\frac{1}{2}$ 时 (2.85) 中第二项要比第一项大的多，且当 $J>40$ 时这条曲线位于图示 B 中相应曲线的上方. 在 5.8 节中将进一步分析解释这些现象.

对全隐式格式 (图示 D) 最大值原理也成立，其图示结果较差但在预料之中：取 $\mu=5$ 时收敛阶为 $O(\Delta t)=O(J^{-2})$; 而取 $\Delta t/\Delta x=\frac{1}{2}$ 时得到误差阶仅为 $O(\Delta t)=O(J^{-1})$.

因为没有考虑到各自的计算量所以这些曲线并没有真实的刻画出各种 θ- 方法的相对效率. 因此在图 2-8 中的曲线均以计算量的一种度量为横轴来描绘同一数值结果：对每种方法计算量应该和总网格点数 $(\Delta x\Delta t)^{-1}$ 大致成比例，辅以显式格式大约需要隐式格式一半计算量的原则. 图示 B 中的两条线不再相同：因为对固定的 ν 随着 J 的增长时间步长 Δt 减小的速度比 μ 固定时慢许多，因此也就需要较少的计算量. 从图中可以看出，对这个问题如果取足够大的 J 而去除初始的振荡，则在所有测试的格式中 $\nu=\frac{1}{2}$ 的 Crank-Nicolson 格式是最有效的；不过与 $\nu=\frac{1}{20}$ 的曲线比较却提示 ν 取其他值时有可能效果会更好.

2.11 最大值原理和 $\mu(1-\theta)\leqslant\frac{1}{2}$ 时的收敛性

如果我们问，在用差分格式逼近 $u_t=u_{xx}$ 时，除了考虑其随 $\Delta t,\Delta x\to 0$ 的收敛性 (以及必要的稳定性和合理的精度阶) 之外还应该考虑何种性质，则下一个自然的要求就是最大值原理. 因为我们知道数学上 (并且从常识来说，例如 u 表示温度) $u(x,t)$ 的上界和下界由初始值和直到时间 t 的边界条件的上下界给出. 在 2.6 节中显式格式收敛性的证明也正是基于这样的一个原理. 而且如果计算结果不具有这种性质，则所有的工程客户都会对此感到非常失望. 我们将此总结成下面这个一般化的定理.

定理 2.2. 由参数满足 $0\leqslant\theta\leqslant 1$ 和 $\mu(1-\theta)\leqslant\frac{1}{2}$ 的 θ-方法(2.75) 产生的 $\{U_j^n\}$ 满足

$$U_{\min}\leqslant U_j^n\leqslant U_{\max}, \tag{2.89}$$

其中

$$U_{\min} := \min\left\{U_0^m,\ 0 \leqslant m \leqslant n;\ U_j^0,\ 0 \leqslant j \leqslant J;\ U_J^m,\ 0 \leqslant m \leqslant n\right\}, \tag{2.90}$$

和

$$U_{\max} := \max\left\{U_0^m,\ 0 \leqslant m \leqslant n;\ U_j^0,\ 0 \leqslant j \leqslant J;\ U_J^m,\ 0 \leqslant m \leqslant n\right\}. \tag{2.91}$$

对最终满足该稳定性条件的任何加密路径，若初始条件和狄利克雷边界条件相容且足够光滑以使得截断误差 $T_j^{n+1/2}$ 沿加密路径在区域上一致趋于零，则 (2.75) 给出的逼近在 $[0,1]\times[0,t_F]$ 上一致收敛.

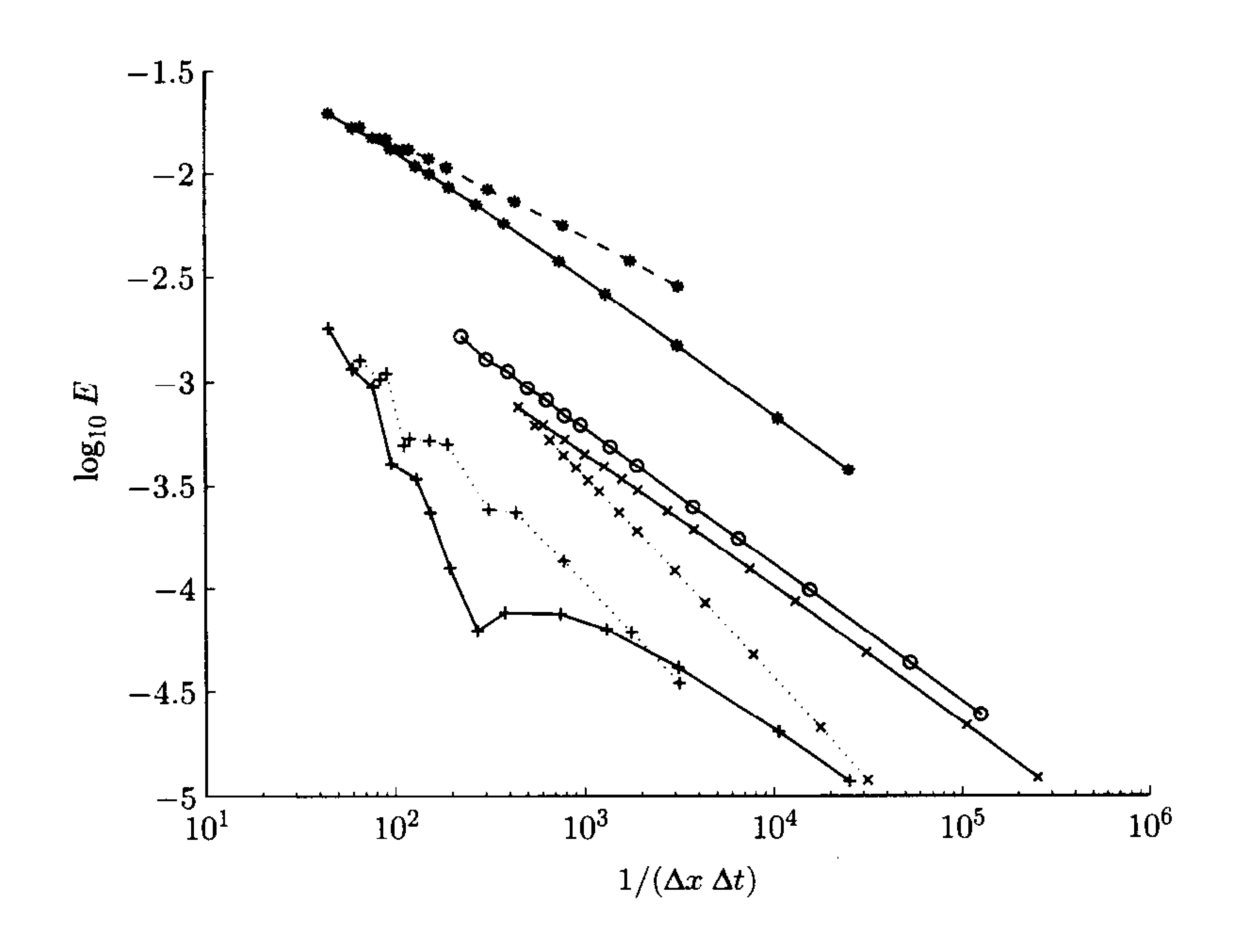


图 2-8 在区域 $[0,1]\times[0.1,1]$ 上，不同格式的最大值误差 E 相对于总网格点数的图

A:	$\theta=0,\ \mu=\frac{1}{2}$	—○—○—○—○
B:	$\theta=\frac{1}{2},\ \mu=\frac{1}{2}$,	—×—×—×—
	$\theta=\frac{1}{2},\ \nu=\frac{1}{20}$	···×···×···
C:	$\theta=\frac{1}{2},\ \mu=5$,	—+—+—+—
	$\theta=\frac{1}{2},\ \nu=\frac{1}{2}$	···+···+···
D:	$\theta=1,\ \mu=5$,	—*—*—*—
	$\theta=1,\ \nu=\frac{1}{2}$	---*---*---*---

证明. 将 (2.75) 改写为以下形式

$$\begin{aligned}(1+2\theta\mu)U_j^{n+1} &= \theta\mu\left(U_{j-1}^{n+1}+U_{j+1}^{n+1}\right)+(1-\theta)\,\mu\left(U_{j-1}^{n}+U_{j+1}^{n}\right)\\ &\quad+\left[1-2(1-\theta)\mu\right]U_j^n. \end{aligned} \tag{2.92}$$

在定理的假设下右端项所有的系数均是非负且和为 $(1+2\theta\mu)$. 现在假设 U 在某个内点达到它的最大值，记该最大值为 U_j^{n+1}, 且令 U^* 是出现在 (2.92) 右端的 U 的五个值之中最大的. 因为系数是非负的，所以有 $U_j^{n+1}\leqslant U^*$; 但是 U_j^{n+1} 已被假定为最大值，故亦有 $U_j^{n+1}\geqslant U^*$, 因此有 $U_j^{n+1}=U^*$. 事实上，最大值必然在 (2.92) 中系数不为零的每个相邻点达到. 在这样的点上进行相同的讨论，则得到一个每一点上均取到最大值的点列直到达到边界点或初始点 [1], 因此最大值必然在边界点或初始点取得. 同理可知最小值也必然在边界点或初始点取得. 定理第一部分证完.

由截断误差的定义 (参见 (2.84)), 微分方程的解满足与 (2.92) 相似的关系式，只是右端多了一个余项 $\Delta tT_j^{n+1/2}$. 所以误差 $e_j^n=U_j^n-u_j^n$ 满足以下关系式

$$\begin{aligned}(1+2\theta\mu)e_j^{n+1} &= \theta\mu\left(e_{j-1}^{n+1}+e_{j+1}^{n+1}\right)+(1-\theta)\,\mu\left(e_{j-1}^{n}+e_{j+1}^{n}\right)\\ &\quad+\left[1-2(1-\theta)\mu\right]e_j^n-\Delta tT_j^{n+1/2}, \end{aligned} \tag{2.93}$$

其中 $j=1,2,\ldots,J-1$, $n=0,1,\ldots$, 而且满足相应的初始条件和边界条件. 首先不妨假设这些条件均为零，因为 $U_j^0=u_j^0$, $U_0^m=u_0^m$ 及 $U_J^m=u_J^m$. 然后，如同 2.6 节，定义

$$E^n:=\max_{0\leqslant j\leqslant J}\left|e_j^n\right|,\quad T^{n+1/2}:=\max_{1\leqslant j\leqslant J-1}\left|T_j^{n+1/2}\right|. \tag{2.94}$$

因为系数非负，由上面的定义得到

$$(1+2\theta\mu)E^{n+1}\leqslant 2\theta\mu E^{n+1}+E^n+\Delta tT^{n+1/2},$$

故有

$$E^{n+1}\leqslant E^n+\Delta tT^{n+1/2}, \tag{2.95}$$

又 $E^0=0$, 于是得到

$$\begin{aligned}E^n &\leqslant \Delta t\sum_0^{n-1}T^{m+1/2}\\ &\leqslant n\Delta t\max_m T^{m+1/2}. \end{aligned} \tag{2.96}$$

在定理的假设条件下，它沿加密路径将趋于零.

到目前为止，我们处理了由有限差分逼近的截断误差所引起的数值误差，而假设边界条件和初始条件是准确使用的. 我们现在假设 U_j^n 的初始条件和边界条件存在误差，用

[1] 原书在该定理证明中多处遗漏了“或初始点”.

ϵ_j^0, ϵ_0^m 和 ϵ_J^m 表示其误差，其中 $0 \leqslant j \leqslant J$ 且 $0 \leqslant m \leqslant N$. 则误差 e_j^n 满足递归关系 (2.93)，且满足初始和边界条件

$$e_j^0 = \epsilon_j^0, \qquad j = 0, 1, \ldots, n,$$
$$e_0^m = \epsilon_0^m, \quad e_J^m = \epsilon_J^m,\ 0 \leqslant m \leqslant N.$$

因此 (依据 Duhamel 原理) e_j^N 可以写为两项和的形式. 第一项满足 (2.93) 和零初始值和边界值；这一项就由 (2.96) 控制. 第二项满足 (2.96) 的齐次形式，即舍弃其中有关 T 的项，且满足给定的非零初始值和边界条件. 根据最大值原理这一项的取值必然在初始值和边界条件的最大和最小值之间. 因此如果初始条件和边界条件是相容的，更确切地说，如果初始值和边界值的误差沿加密路径趋向于零，则数值解的误差沿加密路径也将趋向于零. ■

这个定理的条件，$\mu(1-\theta) \leqslant \frac{1}{2}$ 比在稳定性的傅里叶分析中要求的 $\mu(1-2\theta) \leqslant \frac{1}{2}$ 严格的多；例如，Crank-Nicolson 格式恒满足稳定性条件，但是仅当 $\mu \leqslant 1$ 时才满足最大值原理的要求，而稳定性和收敛性是在最大值原理的基础上在定理中得到证明的. 存在如此大的差距，读者可能会想知道这个定理的条件是否太强. 事实上，最大值原理的条件已经很强了，但它确实稍微有些苛刻：例如当 $J = 2$ 且 $U_0^0 = U_2^0 = 0$, $U_1^0 = 1$ 时，仅在满足给定条件时可以得到非负的 $U_1^1 = 1 - 2(1-\theta)\mu$; 但在实际中一般会取较大的 J 值，而这样则可以稍微放宽条件 (参见习题 2.11). 进而，如果在 $U_0^n = U_J^n = 0$ 条件下想要推出

$$\left|U_j^n\right| \leqslant K \max_{0 \leqslant i \leqslant J} \left|U_i^0\right| \qquad \forall j, n \tag{2.97}$$

且在式中有 $K = 1$, 这正是推出误差界 (2.96) 所需要的性质，则最近得到的结果 [1] 显示：其充分必要条件是 $\mu(1-\theta) \leqslant \frac{1}{4}(2-\theta)/(1-\theta)$, 而对于 Crank-Nicolson 格式这就是 $\mu \leqslant \frac{3}{2}$. 较弱的条件 $\mu(1-2\theta) \leqslant \frac{1}{2}$ 仅仅能够保证增长界中的 K 是有界常数，而稳定性定义 (2.55) 中需要的也正是这样的常数. 实际上，这时可以证明 Crank-Nicolson 格式满足 $K \leqslant 23$!

因此最大值原理分析可以被认为是获得稳定性条件的另外一种途径. 与傅里叶分析方法相比，它的优点是可以容易地推广应用到变系数问题 (见 2.15 节); 但是，正如以上我们已经看到的，由此推出的一般仅仅是稳定性的充分条件.

图 2-9 举例说明了这些要点. 这里的模型问题用 Crank-Nicolson 格式求解. 边界条件为在区域的两个端点取零值. 初始条件 U_j^0 在除了中点以外的各点处均取值为零；而在中点处取值为 1. 这对应在 $\frac{1}{2}$ 处有一个尖峰的函数.

$\mu = 2$ 时最大值原理不成立，从图中可以看出这时计算一个时间步后数值解在中点

[1] Kraaijevanger, J.F.B.M. (1992) Maximum norm contractivity of discretization schemes for the heat equation. *Appl. Numer. Math.* **99**, 475-92.

为负值. 在正常情况下这被认为是不可接受的. $\mu=1$ 时最大值原理成立, 这时数值解的值正如所料都在 0 和 1 之间. 可是, 数值解在计算一个时间步后出现两个尖峰, 分别在中点的两边; 在任意 t 时刻原问题的精确解仅有一个最大值. 这些结果都属于一种非常极端的情况, 并且这种不被接受的现象只持续几个时间步; 其后, 数值解在每种情况下都变的很光滑. 尽管如此, 可以看出如果想模拟解的一些快速变化情形, 我们选取的 μ 就应该比稳定性极限 (stability limit) 再小一些.

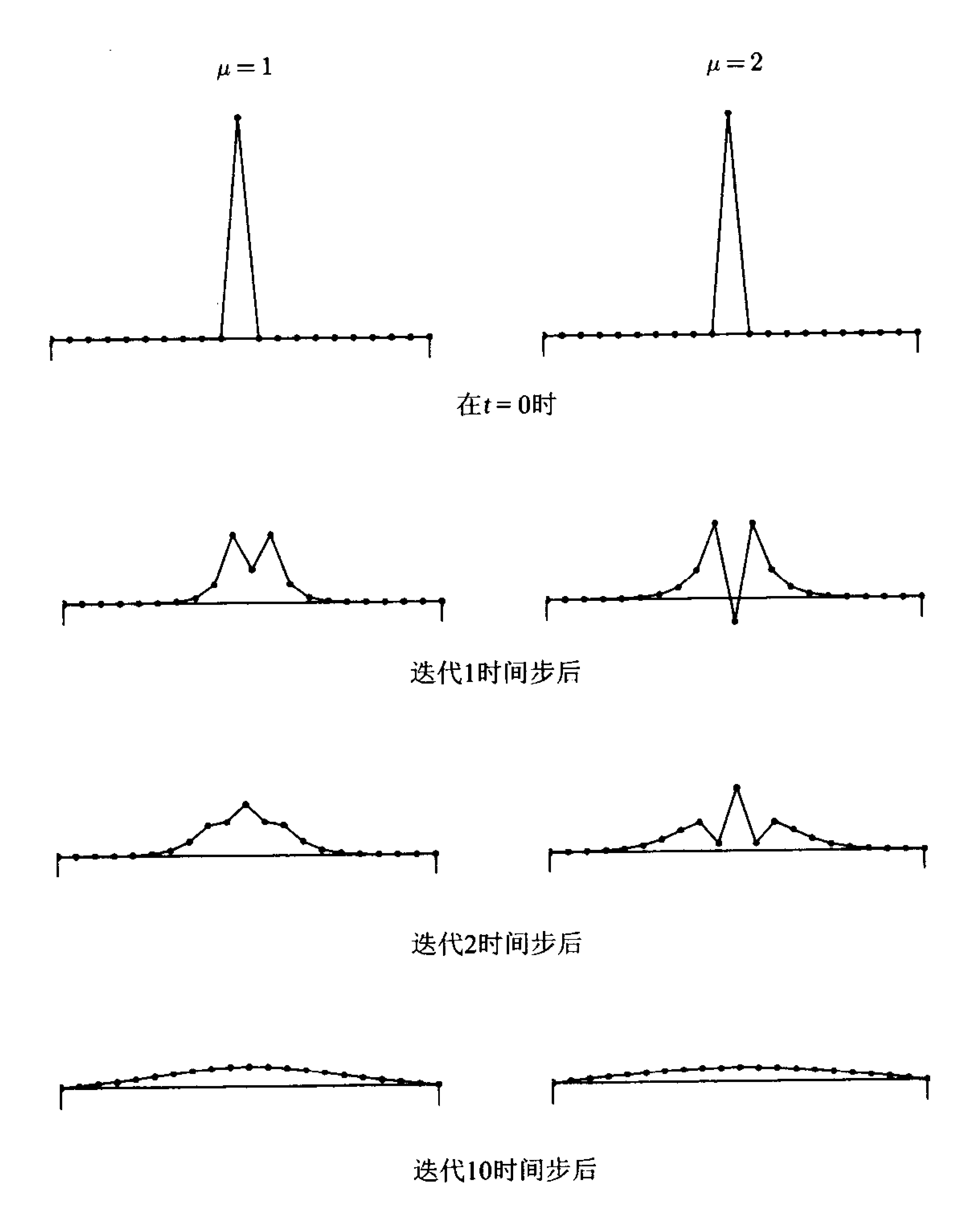


图 2-9 Crank-Nicolson 方法在热传导方程上的应用, 其中初始分布的中点有尖峰; $J=20$, $\Delta x=0.05$

2.12 三时间层格式

我们已经看到利用时间方向的对称性来消去截断误差中偶数次时间导数，使 Crank-Nicolson 格式与显式格式相比提高了精度．但是使用隐式方法也必然增加了额外的计算复杂度．这也提示我们有可能使用超过两个时间层的格式来提高精度并且同时保持显式格式的效率．

例如，用对称中心差分逼近时间导数得到一个显式三层格式

$$\frac{U_j^{n+1}-U_j^{n-1}}{2\Delta t}=\frac{U_{j+1}^n-2U_j^n+U_{j-1}^n}{(\Delta x)^2}. \tag{2.98}$$

容易看出这个逼近的截断误差中仅含有 Δx 和 Δt 的偶数次幂,因此其阶为 $O\left((\Delta x)^2+(\Delta t)^2\right)$. 但是，如果研究这个格式的稳定性，我们会发现具有 (2.72) 形式的一般解需满足

$$\frac{\lambda-1/\lambda}{2\Delta t}=\frac{-4\sin^2\frac{1}{2}k\Delta x}{(\Delta x)^2} \tag{2.99}$$

或

$$\lambda^2+8\lambda\mu\sin^2\frac{1}{2}k\Delta x-1=0. \tag{2.100}$$

这个关于 λ 的二次方程有两个根，故对每个 k 给出两个波型解．两个根都是实的，且其和为负值，其积为 -1. 因此它们中有一个的模大于 1, 它对应于一个不稳定的波型．故这个格式在实际中是无用的，因为 μ 无论取何值它总是不稳定的．

当然这个结果并不意味着每个三层显式格式都是不稳定的．格式

$$\frac{U_j^{n+1}-U_j^{n-1}}{2\Delta t}=\frac{\theta\delta_x^2U_j^n+(1-\theta)\delta_x^2U_j^{n-1}}{(\Delta x)^2} \tag{2.101}$$

当 $\theta\leqslant\frac{1}{2}$ 和 $4(1-\theta)\mu\leqslant 1$ 成立时，其两个波型解都满足 $|\lambda|\leqslant 1$. 我们将证明留作练习 (见练习 2.6); 但是它的稳定性条件和我们的第一个简单格式一样差．

2.13 更一般的边界条件

考虑更一般的模型问题，在 $x=0$ 处引入含导数的边界条件

$$\frac{\partial u}{\partial x}=\alpha(t)u+g(t),\quad \alpha(t)\geqslant 0. \tag{2.102}$$

利用空间向前差分代替空间导数，则可以用

$$\frac{U_1^n-U_0^n}{\Delta x}=\alpha^nU_0^n+g^n \tag{2.103}$$

逼近这个边界条件，并且由此给出边界值 U_0^n 的公式

$$U_0^n = \beta^n U_1^n - \beta^n g^n \Delta x, \tag{2.104a}$$

其中

$$\beta^n = \frac{1}{1+\alpha^n \Delta x}. \tag{2.104b}$$

然后可以和处理狄利克雷边界条件一样应用 θ-方法(2.75). 我们需要求解通常的三对角线性方程组，但是此时这个方程组是关于 J 个未知量 (即新的时间步上的内点和左端边界点的值) 的 J 个方程. (2.104a) 是该方程组的第一个方程，而且显然这个增广系统仍然是三对角的. 因为已假定 $\alpha(t) \geqslant 0$, 因而有 $0 < \beta^n \leqslant 1$, 且除非 $\alpha(t) = 0$, 系数仍然满足条件 (2.67) 和 (2.68). 但当 $\alpha(t) = 0$ 时有 $b_0 = c_0$, $a_0 = 0$ 和 $e_0 = 1$[1].

考虑该格式的精度和稳定性时，需要特别注意第一个内点. 可以用 (2.104a) 来消去 U_0^n; 第一个内点处的二阶差分的形式变为

$$\delta_x^2 U_1^n = U_2^n - (2-\beta^n)U_1^n - \beta^n g^n \Delta x. \tag{2.105}$$

因此根据通常截断误差的定义，经过一些处理后可以证明全局误差满足一个新的关系

$$\begin{aligned}\left[1+\theta\mu(2-\beta^{n+1})\right] e_1^{n+1} \quad = \quad & \left[1-(1-\theta)\mu(2-\beta^n)\right] e_1^n \\ & +\theta\mu e_2^{n+1} + (1-\theta)\mu e_2^n \\ & -\Delta t T_1^{n+1/2}.\end{aligned} \tag{2.106}$$

而不是 (2.93). 这个方程和其他网格点处的方程不同，它使我们无法用傅里叶分析的方法来分析该方程组. 但是可以应用前一节中讨论的最大值原理：首先注意到，如果 $\mu(1-\theta) \leqslant \frac{1}{2}$ 成立，则只要 α^n 非负， (2.106) 中所有的系数就都是非负的；并且如果

$$\theta(1-\beta^{n+1}) \geqslant -(1-\theta)(1-\beta^n), \tag{2.107}$$

则式中右端系数之和不大于左端的系数之和；而当 $\alpha(t) \geqslant 0$ 时上式恒成立. 因此我们可以像以前一样推导出用截断误差表示的全局误差界的关系式 (2.96). 在这些讨论中假设条件 $\alpha(t) \geqslant 0$ 的重要性是明显的：相反的假设对应于和表面温度成正比的热量流入而不是流出，这将导致按照指数规律增长的解. 在实际问题中这是不可能发生的，且在任何情况下都将导致一个不适定的问题.

剩下来要估计的是截断误差 $T_1^{n+1/2}$. 我们仅限于考虑显式格式 $\theta = 0$ 的情形，我们在第一个内点处将其展开，假设直接把 (2.103) 看作 $(0, t_n)$ 点处的边界条件，为得到精确解，在这点展开，得到

$$\frac{u_1^n - u_0^n}{\Delta x} - \alpha^n u_0^n - g^n = \left[\frac{1}{2}\Delta x u_{xx} + \frac{1}{6}(\Delta x)^2 u_{xxx} + \cdots\right]_0^n. \tag{2.108}$$

---

[1]原书此处为： $b_0 = a_0$ 和 $e_0 = 1$.

为了消去 u_0^n 项，把逼近边界条件的 (2.103) 乘以一个合适的乘子加到差分方程中，便得到如下形式的截断误差

$$\begin{aligned} T_1^{n+1/2} &= \frac{u_1^{n+1}-u_1^n}{\Delta x} - \frac{\delta_x^2 u_1^n}{(\Delta x)^2} - \frac{\beta^n}{\Delta x}\left[\frac{u_1^n-u_0^n}{\Delta x} - \alpha^n u_0^n - g^n\right] \\ &= \left[\frac{1}{2}\Delta t u_{tt} - \frac{1}{12}(\Delta x)^2 u_{xxxx} + \cdots\right]_1^n - \beta^n\left[\frac{1}{2}u_{xx} + \cdots\right]_0^n, \end{aligned}$$

故有

$$T_1^{n+1/2} \approx -\frac{1}{2}\beta^n u_{xx}. \tag{2.109}$$

随网格尺度趋于零上式并不趋向于零，虽然可以通过更细致的分析得到收敛性，但这里我们不打算这样做.

然而，一个细小的修改可以修正这个问题，选择新的网格点，它们仍然是等间距的，但是边界点 $x=0$ 位于开始两个网格点的中点. 另一边的边界 $x=1$ 仍然如前是最后一个网格点. 现在用更精确的方式逼近边界条件

$$\frac{U_1^n-U_0^n}{\Delta x} = \frac{1}{2}\alpha^n(U_0^n+U_1^n)+g^n, \tag{2.110}$$

$$U_0^n = \frac{1-\frac{1}{2}\alpha^n\Delta x}{1+\frac{1}{2}\alpha^n\Delta x}U_1^n - \frac{\Delta x}{1+\frac{1}{2}\alpha^n\Delta x}g^n. \tag{2.111}$$

然后将 (2.108) 替换为在 $j=\frac{1}{2}$ 点处的展式，得到

$$\begin{aligned} &\frac{u_1^n-u_0^n}{\Delta x} - \frac{1}{2}\alpha^n(u_0^n+u_1^n) - g^n = \\ &\qquad \left[\frac{1}{24}(\Delta x)^2 u_{xxx} - \frac{1}{8}\alpha^n(\Delta x)^2 u_{xx} + \cdots\right]_{1/2}^n, \end{aligned} \tag{2.112}$$

于是有

$$\begin{aligned} T^{n+1/2} &= \left[\frac{1}{2}\Delta t u_{tt} - \frac{1}{12}(\Delta x)^2 u_{xxxx} + \cdots\right]_1^n \\ &\quad - \frac{1}{1+\frac{1}{2}\alpha^n\Delta x}\left[\frac{1}{24}\Delta x(u_{xxx} - 3\alpha^n u_{xx}) + \cdots\right]_0^n \\ &= O(\Delta x). \end{aligned} \tag{2.113}$$

只要把 (2.106) 稍做修改，就可以证明收敛性. 其实，在第 6 章将给出一个基于最大值原理的更精确的误差分析：尽管边界附近有 $O(\Delta x)$ 阶的截断误差，整体误差仍是 $O\left((\Delta x)^2\right)$ 阶的.

另一种应用更广泛的方法是仍保持 $x=0$ 是第一个网格点，但是在区域外引入一个

虚拟的点值 U_{-1}^n 以使用中心差分格式得到

$$\frac{U_1^n - U_{-1}^n}{2\Delta x} = \alpha^n U_0^n + g^n. \tag{2.114}$$

在 $x = 0$ 点仍然可以用通常的差分逼近，故可以消去 U_{-1}^n. 对 θ-方法, 我们有

$$\begin{aligned}\frac{U_0^{n+1} - U_0^n}{\Delta t} &- \frac{\delta_x^2}{(\Delta x)^2}\left[\theta U_0^{n+1} + (1-\theta)U_0^n\right] \\ &- \frac{2\theta}{\Delta x}\left[\frac{U_1^{n+1} - U_{-1}^{n+1}}{2\Delta x} - \alpha^{n+1}U_0^{n+1} - g^{n+1}\right] \\ &- \frac{2(1-\theta)}{\Delta x}\left[\frac{U_1^n - U_{-1}^n}{2\Delta x} - \alpha^n U_0^n - g^n\right] = 0. \end{aligned} \tag{2.115}$$

显然，要计算截断误差需要把

$$\frac{2\theta}{\Delta x}\left[\frac{u_1^{n+1} - u_{-1}^{n+1}}{2\Delta x} - \alpha^{n+1}u_0^{n+1} - g^{n+1}\right] = \theta\left[\frac{1}{3}\Delta x u_{xxx}\right]_0^{n+1} + \cdots \tag{2.116}$$

加入到一般的截断误差项中. 如果我们把 (2.115) 整理为

$$\begin{aligned}\left[1 + 2\theta\mu(1 + \alpha^{n+1}\Delta x)\right]U_0^{n+1} = &\left[1 - 2(1-\theta)\mu(1 + \alpha^n\Delta x)\right]U_0^n \\ &+ 2\theta\mu U_1^{n+1} \\ &- 2\mu\Delta x\left[\theta g^{n+1} + (1-\theta)g^n\right],\end{aligned} \tag{2.117}$$

并且把定理 2.2 中条件略微加强为

$$\mu(1-\theta)(1 + \alpha^n\Delta x) \leqslant \frac{1}{2}. \tag{2.118}$$

则基于最大值原理的误差分析仍然成立.

这一节中我们考虑的热传导方程问题在左端处是导数边界条件，而另一端是狄利克雷边界条件. 相同的方法对右端是导数边界条件或者在两端都是的情况也是适用的. 经过与前面类似的分析就会发现，在 $x = 1$ 处导数边界条件的形式应为

$$\frac{\partial u}{\partial x} = \beta(t)u + g(t), \ \beta(t) \leqslant 0. \tag{2.119}$$

为了演示处理边界条件的不同方法，在 $0 < x < 1$ 上计算问题 $u_t = u_{xx}$ 的解，其初始条件为 $u(x, 0) = 1 - x^2$, 并在左端点满足诺伊曼边界条件 $u_x(0, t) = 0$, 右端点满足狄利克雷边界条件 $u(1, t) = 0$. 使用 Crank-Nicolson 格式求解，取 $J = 10$, $\mu = 1$, 即有 $\Delta t = 0.01$. 对以上给出的三种方法进行分析：即用向前差分格式逼近 $u_{x,t}$ 方法，结合虚拟点 U_{-1}^n 使用中心差分的方法，以及构造网格使边界点在开始两网格点的中点处的方法，图 2-10 显示了相应数值解的最大误差，作为 t_n 的函数. 第二和第三种方法的数值结果非常相似，但是与第一种方法却有显著的不同；在这个例子中第一种方法的误差大约是后两者的 50 倍.

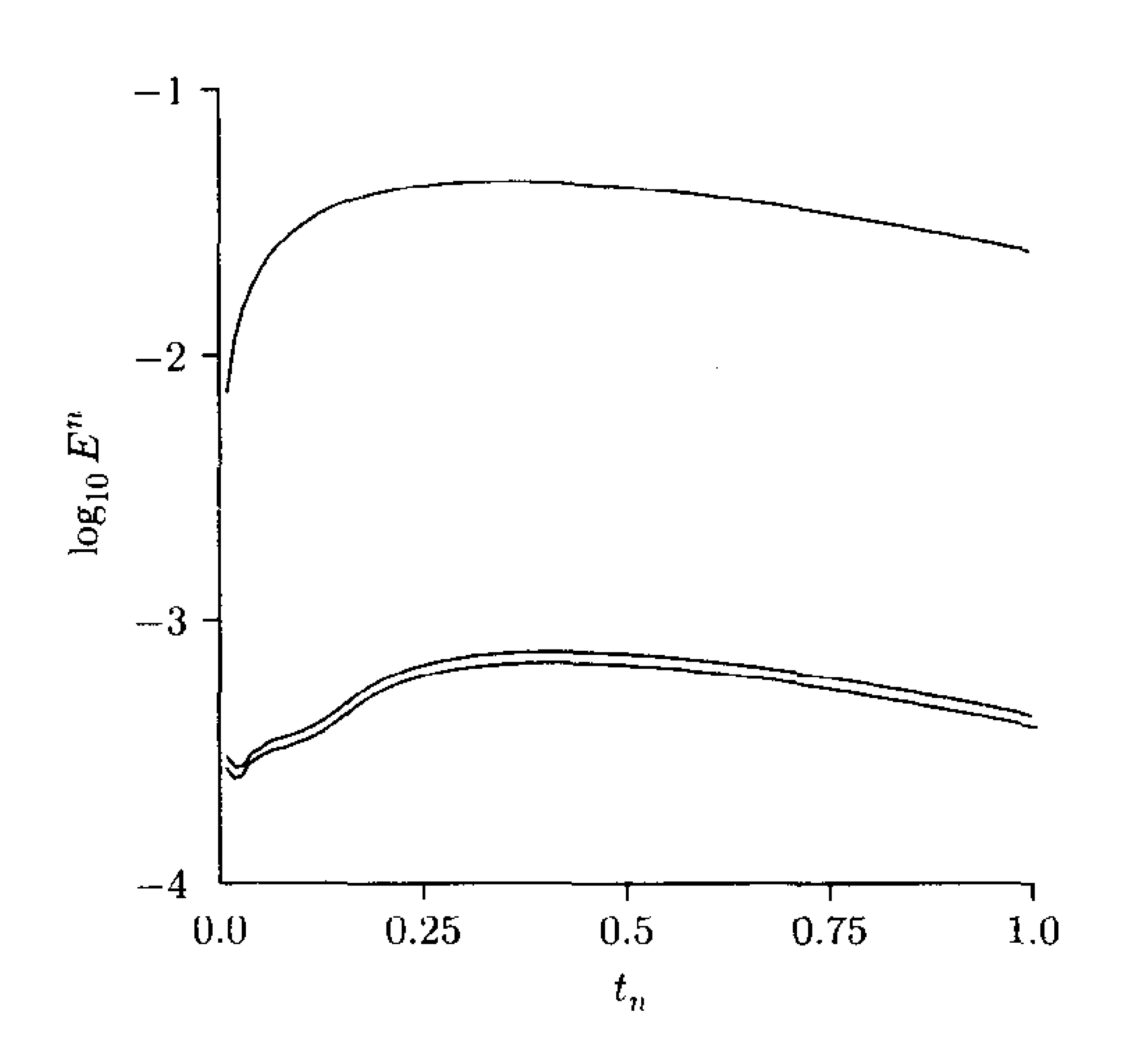


图 2-10　诺伊曼边界条件的近似对取 $J = 10, \Delta x = 0.1$ 的 Crank-Nicolson 格式误差的影响；顶部曲线使用 (2.103)，下面的两条分别使用 (2.114) 和 (2.110)

同时注意，与图 2-4 中两端都为狄利克雷边界条件的情况误差性态相对比，在一定范围内每种方法的最大值误差都随着 n 增长，过一段之后才很缓慢地开始下降. 下一节我们将考虑两端都是诺伊曼边界条件的情况.

2.14　热量守恒性质

在模型问题热传导方程 $u_t = u_{xx}$ 中定义 t 时刻的系统总热能

$$h(t) = \int_0^1 u(x,t)\mathrm{d}x. \tag{2.120}$$

由微分方程，有

$$\frac{\mathrm{d}h}{\mathrm{d}t} = \int_0^1 u_t \mathrm{d}x = \int_0^1 u_{xx}\mathrm{d}x = [u_x]_0^1. \tag{2.121}$$

在狄利克雷边界条件的时候这样做并没有什么帮助，但是假定在两个端点处都是诺伊曼边界条件，即 $u_x(0,t) = g_0(t)$ 和 $u_x(1,t) = g_1(t)$. 继而得到

$$\frac{\mathrm{d}h}{\mathrm{d}t} = g_1(t) - g_0(t), \tag{2.122}$$

对这个常微分方程求积分则可以求得 h .

现在对 θ- 方法的方程 (2.75) 做类似的处理. 对 (2.75) 成立的所有点求和以引入离散

总热能：

$$H^n = \sum_{1}^{J-1} \Delta x U_j^n. \tag{2.123}$$

然后，利用有限差分记号 $\delta_x^2 U_j = \Delta_{+x} U_j - \Delta_{+x} U_{j-1}$，得到

$$\begin{aligned} H^{n+1} - H^n &= \frac{\Delta t}{\Delta x} \sum_{1}^{J-1} \delta_x^2 \left[\theta U_j^{n+1} + (1-\theta) U_j^n\right] \\ &= \frac{\Delta t}{\Delta x} \left\{\Delta_{+x} \left[\theta U_{J-1}^{n+1} + (1-\theta) U_{J-1}^n\right]\right. \\ &\qquad \left. - \Delta_{+x} \left[\theta U_0^{n+1} + (1-\theta) U_0^n\right]\right\}. \end{aligned} \tag{2.124}$$

余下的分析将依赖于如何逼近边界条件．考虑最简单的逼近，如 (2.103)：即令 $U_1^n - U_0^n = \Delta x g_0^n$, $U_J^n - U_{J-1}^n = \Delta x g_1^n$. 于是得到作为 (2.122) 的逼近式：

$$H^{n+1} - H^n = \Delta t \left[\theta (g_1^{n+1} - g_0^{n+1}) + (1-\theta)(g_1^n - g_0^n)\right]; \tag{2.125}$$

这个逼近可能是非常精确的，尽管如我们所知， U^n 可能并不是 u^n 的一个好的点值逼近．特别地，如果 g_0 和 g_1 都和 t 无关，则一个时间步中 H 的变化量与 $h(t)$ 的变化量严格相等．这如何解释呢？

显然，为了更好地理解这种匹配得很好的对应关系，我们应该将 (2.123) 与 (2.120) 尽可能地紧密联系起来．如果 U 和 u 均是常数，这种联系意味着应该取 $(J-1)\Delta x = 1$，而不是像我们所假设的 $J\Delta x = 1$; 同时应该把 U_j^n 与

$$u_j^n := \frac{1}{\Delta x} \int_{(j-1)\Delta x}^{j\Delta x} u(x, t_n) \mathrm{d}x, \ j = 1, 2, \ldots, J-1, \tag{2.126}$$

相比较，因此它应该表示位于区间中心 $(j - \frac{1}{2})\Delta x$ 处的值；我们还有

$$h(t_n) = \sum_{1}^{J-1} \Delta x \ u_j^n. \tag{2.127}$$

注意对数值解的这种解释与我们在分析截断误差时得到的格式 (2.110)~(2.113) 非常匹配．这也意味着初始条件应该取为 $U_j^0 = u_j^0$, 其中 u_j^0 由 (2.126) 式定义．这样处理之后，若边界条件与时间无关，则对所有的 n 均有 $H^n = h(t_n)$. 此外，不难看出对任意常数 C，函数

$$\hat{u}(x, t) := (g_1 - g_0)t + \frac{1}{2}(g_1 - g_0)x^2 + g_0 x + C \tag{2.128}$$

满足微分方程和两个边界条件．可以证明，对任意给定的初始条件，问题的精确解将随 $t \to \infty$ 趋于如上形式的解．因为函数 (2.128) 是 t 的线性函数且关于 x 是二次的，所以它也必然严格满足有限差分方程；若用以上方式解释数值解，则随 t 的增长数值解产生的

误差将趋于零. 正如前面所看到的, 处理诺伊曼边界条件问题时若用通常的方式解释误差, 则一般差分逼近所产生的误差就可能会在开始阶段随 n 增长, 然后仅以非常缓慢的速率下降.

我们通过一个齐次热传导方程的解来演示这些现象. 方程在 $x=0$ 和 $x=1$ 处赋诺伊曼边界条件 $u_x=0$, 且初始条件为 $u(x,0)=1-x^2$. 图 2-11 给出了三种情形下最大值误差作为 t_n 的函数的曲线, 它们均使用 $\mu=\frac{1}{2}$ 的 Crank-Nicolson 方法, 且取 $J=10$. 顶

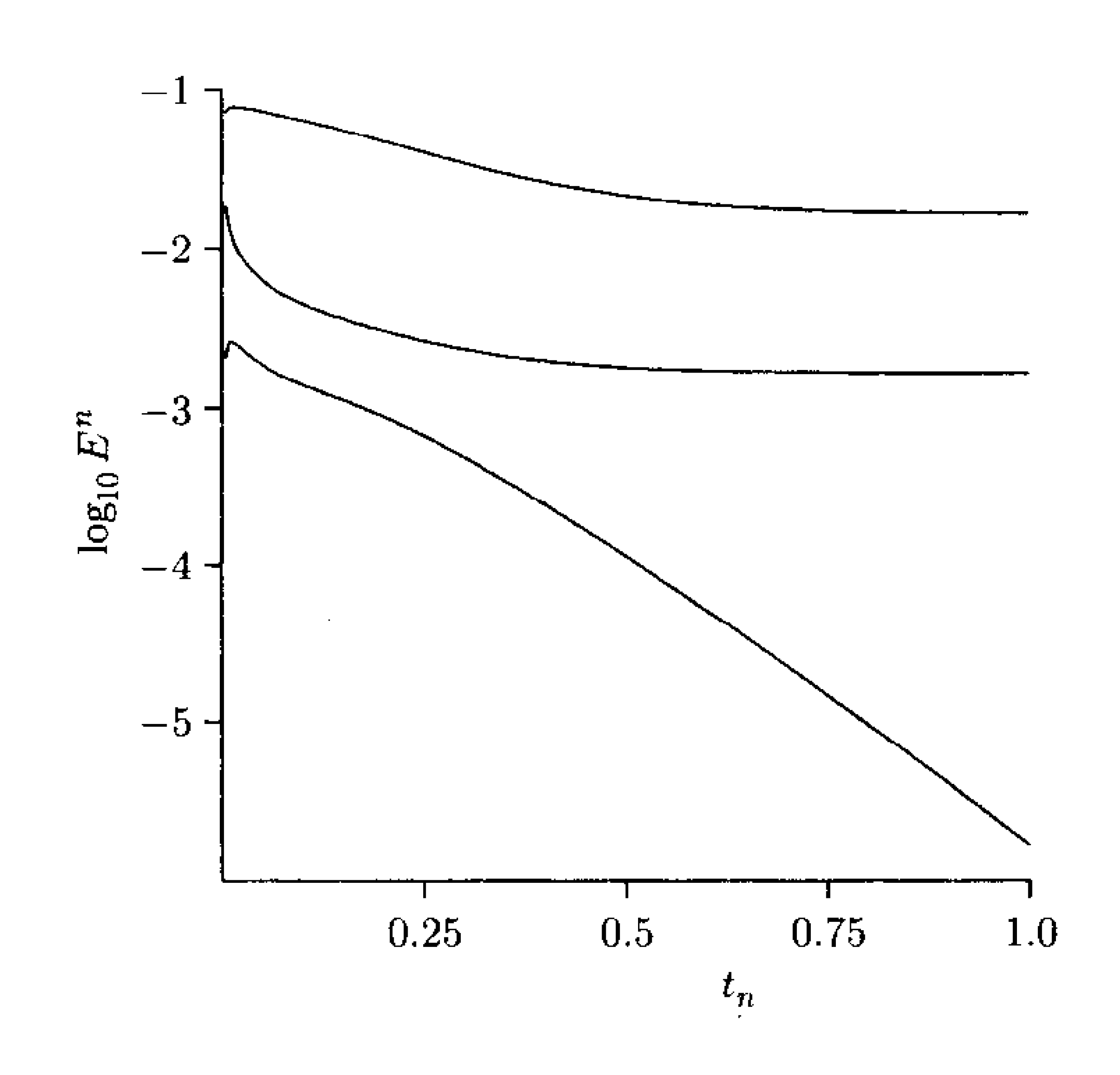


图 2-11　对纯诺伊曼问题误差解释方式的效果: 顶部曲线对应边界条件 (2.103), 而第二条对应 (2.114), 二者都采用通常的方式取初始值和定义 E^n; 底部曲线和顶部曲线采用相同的计算方法, 但是其初始条件使用 (2.126), 且用热量守恒原理重新解释了误差

部曲线对应的是用 (2.103) 逼近诺伊曼边界条件且用通常的方式解释误差 E^n 的结果, 其 $\Delta x=0.1$, $\Delta t=0.005$. 第二条曲线除了改用 (2.114) 逼近边界条件之外和第一条曲线采用方法完全一样. 明显地, 二者的残差都很大. 然而, 底部曲线虽然和顶部曲线使用相同的方法, 但是初始数据用 (2.126) 中得到的 u_j^n 给出, 并且重新解释误差以反映热量守恒原理: 即 $E^n:=\max\left\{|U_j^n-u_j^n|,\ j=0,1,2,\ldots,J\right\}$; 且 $t_n=n\Delta t=n\mu(\Delta x)^2=\frac{1}{2}n(J-1)^{-2}$. 从图中我们可以清楚地看到如上讨论所预期的误差下降情形.

2.15　更一般的线性问题

到目前为止我们所考虑的热传导方程对应的物理问题都满足其物理性质关于时间是常数的条件, 且独立于 x. 一般地, 这些性质是 x, 或 t, 或二者的函数. 特别地, 依赖 x 的情形通常是模拟近乎一维的细棒中的热流, 其横截面面积依赖于 x. 因此, 我们来简单地

研究一下如何将目前讨论的方法做适当的改动以便应用于更一般的问题.

首先，考虑问题

$$\frac{\partial u}{\partial t} = b(x,t)\frac{\partial^2 u}{\partial x^2}, \tag{2.129}$$

其中函数 $b(x,t)$ 通常假定是严格正的. 简单地推广显式格式 (2.19), 得到

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{(\Delta x)^2} b_j^n \left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right), \tag{2.130}$$

其中 $b_j^n = b(x_j, t_n)$. 这个格式的实现和先前的显式格式一样简单，并且误差分析也几乎不需要更改. 如同处理 (2.19), 同样泰勒级数展开得到表达式

$$T(x,t) = \frac{1}{2}\Delta t u_{tt} - \frac{1}{12}b(x,t)(\Delta x)^2 u_{xxxx} + \cdots. \tag{2.131}$$

虽然推出 (2.44) 式的分析仍然有效，但是稳定性条件必须替换为：对区域中任意的 x 和 t 均满足

$$\frac{\Delta t}{(\Delta x)^2} b(x,t) \leqslant \frac{1}{2}. \tag{2.132}$$

最终的误差界变为

$$E^n \leqslant \frac{1}{2}\Delta t \left[M_{tt} + \frac{B(\Delta x)^2}{6\Delta t} M_{xxxx}\right] t_F \tag{2.133}$$

其中 B 是在区域 $[0,1]\times[0,t_F]$ 上 $b(x,t)$ 的一致上界.

通过几个略微不同的修改方式，θ- 方法可以用于处理这种更一般的问题. 方程 (2.75) 显然可以推广为

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{(\Delta x)^2} b^* \left[\theta\delta_x^2 U_j^{n+1} + (1-\theta)\delta_x^2 U_j^n\right], \tag{2.134}$$

但是如何最优地选取 b^* 的值却不是显而易见的. 在先前分析这个格式的截断误差的过程中，我们在中点 $(x_j, t_{n+\frac{1}{2}})$ 处做泰勒级数展开. 这建议选择

$$b^* = b_j^{n+\frac{1}{2}}; \tag{2.135}$$

事实上，易见基于这个选择除了 (2.84) 中包含了一个额外的因子 b 之外，先前的截断误差表达式并没有其他改动，它变为

$$\begin{aligned} T_j^{n+1/2} = &\Big[(\frac{1}{2}-\theta)\Delta t\, u_{xxt} - \frac{b}{12}(\Delta x)^2 u_{xxxx} + \frac{1}{24}(\Delta t)^2 u_{ttt} \\ &-\frac{b}{8}(\Delta t)^2 u_{xxtt} + \frac{1}{12}(\frac{1}{2}-\theta)\Delta t(\Delta x)^2 u_{xxxxt} \\ &-\frac{2b}{6!}(\Delta x)^4 u_{xxxxxx} + \cdots\Big]_j^{n+1/2}. \end{aligned} \tag{2.136}$$

利用最大值原理，收敛性证明也没有改变，只是稳定性条件变为：在区域中的所有点处都满足

$$\frac{\Delta t}{(\Delta x)^2}(1-\theta)b(x,t) \leqslant \frac{1}{2}. \tag{2.137}$$

这样选择 b^* 要求在时间的半步长处计算 $b(x,t)$, 这在一些问题中是不易实现的. 另一种明显的选择方案是

$$b^* = \frac{1}{2}(b_j^{n+1} + b_j^n). \tag{2.138}$$

在中点处做泰勒展开，得到

$$b^* = \left[b + \frac{1}{4}(\Delta t)^2 b_{tt} + \cdots\right]_j^{n+1/2}. \tag{2.139}$$

这将在截断误差的展式中引入额外的一个关于 b_{tt} 的高阶项.

线性抛物型方程最一般的形式是

$$\frac{\partial u}{\partial t} = b(x,t)\frac{\partial^2 u}{\partial x^2} - a(x,t)\frac{\partial u}{\partial x} + c(x,t)u + d(x,t), \tag{2.140}$$

如前假设 $b(x,t)$ 总是正的. 这里符号的选择是为了和后面的章节，特别是与 5.7 节中的 (5.48) 式相一致. 特别地， $a(x,t)$ 前面的负号是为了方便但是并不重要，因为 $a(x,t)$ 可取任意符号；仅要求 $b(x,t)$ 是正的. 我们可以很容易地为这个方程构造一个显式格式；只有 $\partial u/\partial x$ 需要新的考虑. 因为逼近二阶导数用的是中心差分，自然地也用中心差分逼近一阶导数，从而得到格式

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{b_j^n}{(\Delta x)^2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) - \frac{a_j^n}{2\Delta x}\left(U_{j+1}^n - U_{j-1}^n\right) + c_j^n U_j^n + d_j^n. \tag{2.141}$$

计算截断误差的主项是容易的，我们把它留作练习. 然而，在分析误差 e_j^n 性态的过程中产生了新的困难. 就像由对简单问题的分析得到 (2.42) 一样，这里我们得到

$$\begin{aligned} e_j^{n+1} &= e_j^n + \mu_j^n\left(e_{j+1}^n - 2e_j^n + e_{j-1}^n\right) - \frac{1}{2}\nu_j^n\left(e_{j+1}^n - e_{j-1}^n\right) + \Delta t\, c_j^n e_j^n - \Delta t T_j^n \\ &= \left(1 - 2\mu_j^n + \Delta t\, c_j^n\right)e_j^n + (\mu_j^n - \frac{1}{2}\nu_j^n)e_{j+1}^n + (\mu_j^n + \frac{1}{2}\nu_j^n)e_{j-1}^n - \Delta t T_j^n, \end{aligned} \tag{2.142}$$

其中

$$\mu_j^n = \frac{\Delta t}{(\Delta x)^2}b_j^n, \quad \nu_j^n = \frac{\Delta t}{\Delta x}a_j^n. \tag{2.143}$$

为了得到和前面相似的关于 e_j^n 的界，需要保证方程右端涉及 e^n 的三项的系数都是非负的，而且其和不大于一. 我们总是假设函数 $b(x,t)$ 是严格正的，但是一般不能对 $a(x,t)$ 的

符号做任何假设．因此，除了 $c_j^n \leqslant 0$ 之外，还需要条件

$$\frac{1}{2}\left|\nu_j^n\right| \leqslant \mu_j^n, \tag{2.144}$$

$$2\mu_j^n - \Delta t c_j^n \leqslant 1. \tag{2.145}$$

因为有条件 $c_j^n \leqslant 0$，故第二个条件要比简单情形时稍微严格一点；确实，如果条件 $0 \leqslant c(x,t) \leqslant C$ 成立，则条件 (2.145) 代表稍稍放松 μ 的约束，但那样仅可以得到 $E^{n+1} \leqslant (1+C\Delta t)E^n + T\Delta t$. 然而，第一个条件要严格得多．如果，我们用其 Δx 和 Δt 的表达式代替 ν 和 μ, 则该条件表示为

$$\Delta x \leqslant \frac{2b_j^n}{|a_j^n|}, \quad 或 \quad \frac{|a_j^n|\Delta x}{b_j^n} \leqslant 2, \tag{2.146}$$

并且上式对任意的 n 和 j 均须成立．因此，Δx 的尺度就受到了限制，这也就意味着 Δt 的尺度受到了限制．

在许多实际问题中函数 $b(x,t)$ 相对于 $a(x,t)$ 来说可能是非常小的．例如，在大部分流体问题中粘性往往很小，上述情况就会发生．在此情况下，一个重要的无量纲参数是 *Péclet*数(Péclet number) UL/ν, 其中 U 是速度，L 是长度而 ν 是粘性．这样的问题和**奇异摄动问题** (singular perturbation problem) 是很相近的，且不能简单地用这种显式的中心差分方法求解：因为由 (2.146), 网格*Péclet*数(mesh Péclet number) 不能大于 2, 其中长度就是网格尺度．例如，如果有[1] $b=1$, $a=1000$, $c=0$, 则条件要求 $\Delta x \leqslant 0.002$, 因而要求 $\Delta t \leqslant 0.000002$. 这样在 x 方向我们至少需要 500 个网格点，而若要计算到一个合理的时间 t_F 则需要巨大的时间步数．

一个简单的避免这个问题的方法是使用向前或向后差分来逼近一阶导数项，而不是用中心差分．例如，若已知 $a(x,t) \geqslant 0$ 和 $c(x,t)=0$, 则可以使用向后差分，差分格式变为

$$\frac{U_j^{n+1}-U_j^n}{\Delta t} = \frac{b_j^n}{(\Delta x)^2}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) - \frac{a_j^n}{\Delta x}\left(U_j^n - U_{j-1}^n\right) + c_j^n U_j^n + d_j^n, \tag{2.147}$$

进而得到

$$\begin{aligned} e_j^{n+1} &= e_j^n + \mu_j^n\left(e_{j+1}^n - 2e_j^n + e_{j-1}^n\right) - \nu_j^n\left(e_j^n - e_{j-1}^n\right) - \Delta t T_j^n \\ &= \left(1 - 2\mu_j^n - \nu_j^n\right)e_j^n + \mu_j^n e_{j+1}^n + \left(\mu_j^n + \nu_j^n\right)e_{j-1}^n - \Delta t T_j^n. \end{aligned} \tag{2.148}$$

为了保证该方程右端所有项的系数均非负，现在只要求满足

$$2\mu_j^n + \nu_j^n \leqslant 1. \tag{2.149}$$

当 $a \neq 0$ 时这个条件对时间步尺度的要求更严格，但是对 Δx 的尺度却没有约束．

如果函数 $a(x,t)$ 变号，则当 a 为正时应该使用向后差分，而当它为负时则应使用向

[1] 原书为：$b=0.001$, $a=1$. 但这样得到 $\Delta t \leqslant 0.002$ 而不是 $\leqslant 0.000002$.

前差分；这种方法被称为*迎风差分* (*upwind differencing*). 不幸地是我们必须以更严格的约束条件为代价才可以保证最大值原理成立. 现在截断误差的阶下降了：向前差分引入的是 Δx 阶的误差，而非中心差分的 $(\Delta x)^2$ 阶. 我们将在讨论双曲型方程的章节中讨论这个问题.

抛物型方程一般也时常以自共轭的形式出现

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(p(x,t)\frac{\partial u}{\partial x}\right), \tag{2.150}$$

这里通常假定函数 $p(x,t)$ 是严格正的. 也可以把这类方程改写为刚才讨论过的形式，如

$$\frac{\partial u}{\partial t}=p\frac{\partial^2 u}{\partial x^2}+\frac{\partial p}{\partial x}\frac{\partial u}{\partial x}. \tag{2.151}$$

不过，直接对方程的原始形式构造差分逼近效果一般更好. 我们使用逼近

$$\left[p\frac{\partial u}{\partial x}\right]^n_{j+1/2}\approx p^n_{j+1/2}\left(\frac{u^n_{j+1}-u^n_j}{\Delta x}\right), \tag{2.152}$$

将其中的 j 全部替换为 $j-1$ 可以得到一个类似的逼近式. 把二者相减并除以 Δx, 便得到方程右端项的一个逼近. 这给出了一个显式差分格式

$$\frac{U^{n+1}_j-U^n_j}{\Delta t}=\frac{1}{(\Delta x)^2}\left[p^n_{j+1/2}\left(U^n_{j+1}-U^n_j\right)-p^n_{j-1/2}\left(U^n_j-U^n_{j-1}\right)\right]. \tag{2.153}$$

记

$$\mu'=\frac{\Delta t}{(\Delta x)^2},$$

则得显示表达式

$$U^{n+1}_j=\left(1-\mu'(p^n_{j+1/2}+p^n_{j-1/2})\right)U^n_j+\mu' p^n_{j+1/2}U^n_{j+1}+\mu' p^n_{j-1/2}U^n_{j-1}. \tag{2.154}$$

从上式可以看出我们先前使用的误差分析方法在这里也同样适用，若

$$\mu' P\leqslant\frac{1}{2}, \tag{2.155}$$

则右端项的系数均是非负的，其中 P 是函数 $p(x,t)$ 在区域中的上界. 因此该格式只对时间步提出了某种约束条件，这是我们预料之中的，而对空间步长 Δx 的尺度则没有加以任何限制.

从同种类型的差分逼近格式出发也容易得到 θ- 方法的推广. 其中细节，包括截断误差主项等的计算，均留作习题 (参见习题 2.7 和 2.8).

2.16 极坐标

一维问题往往是从具有柱对称或球对称性质的三维物理问题得来. 在极坐标下，简

单的热传导方程的形式为

$$\frac{\partial u}{\partial t}=\frac{1}{r^{\alpha}}\frac{\partial}{\partial r}\left(r^{\alpha}\frac{\partial u}{\partial r}\right) \tag{2.156}$$

或

$$u_t=u_{rr}+\frac{\alpha}{r}u_r, \tag{2.157}$$

其中 $\alpha=0$ 时对应的是我们一直考虑的平面对称问题，$\alpha=1$ 对应柱对称，而 $\alpha=2$ 则对应球对称．从形式 (2.156) 或者 (2.157) 入手，刚刚得到的方法可以容易地应用于这些方程．在这个特例下，比较两种形式的稳定性条件发现，它们并没有多少差别．然而在两种情况中，极点 $r=0$ 处都明显存在问题．

无论在二维或者三维中考虑问题的对称性可以得到在极点处 $\partial u/\partial r=0$; 若 u_r 非零，则由 (2.157) 可知在 $r=0$ 处 u_{rr} 为无穷或者 u_t 为无穷，或者二者均为无穷．现在视 t 为常数，把 u 仅看作 r 的函数，并且在 $r=0$ 处泰勒级数展开，得到

$$\begin{aligned}u(r) &= u(0)+ru_r(0)+\frac{1}{2}r^2u_{rr}(0)+\cdots\\ &= u(0)+\frac{1}{2}r^2u_{rr}(0)+\cdots\end{aligned} \tag{2.158}$$

和

$$\begin{aligned}\frac{1}{r^{\alpha}}\frac{\partial}{\partial r}\left(r^{\alpha}\frac{\partial u}{\partial r}\right) &= \frac{1}{r^{\alpha}}\frac{\partial}{\partial r}\left[r^{\alpha}u_r(0)+r^{\alpha+1}u_{rr}(0)+\cdots\right]\\ &= \frac{1}{r^{\alpha}}\left[(\alpha+1)r^{\alpha}u_{rr}(0)+\cdots\right]\\ &= (\alpha+1)u_{rr}(0)+\cdots.\end{aligned} \tag{2.159}$$

在 (2.158) 中用 Δr 代替 r, 得到

$$u(\Delta r)-u(0)=\frac{1}{2}(\Delta r)^2u_{rr}(0)+\cdots, \tag{2.160}$$

并且由此得到区域左端点处的一个差分逼近

$$\frac{U_0^{n+1}-U_0^n}{\Delta t}=\frac{2(\alpha+1)}{(\Delta r)^2}\left(U_1^n-U_0^n\right). \tag{2.161}$$

这种做法同样也适用于 θ-方法.

另一种更物理化的观点直接来源于 (2.156). 如图 2-12,

考虑环形区域的 $r=r_{j-1/2}$ 和 $r=r_{j+1/2}$ 的两个表面之间的热平衡： (2.156) 的右端项 $r^{\alpha}\partial u/\partial r$ 正比于热通量与表面积的积；而通过半径分别为 $r_{j-1/2}$ 和 $r_{j+1/2}$ 的两个表面的热通量之差来提高其间物质的温度，这部分物质的体积正比于 $(r_{j+1/2}^{\alpha+1}-r_{j-1/2}^{\alpha+1})/(\alpha+1)$.

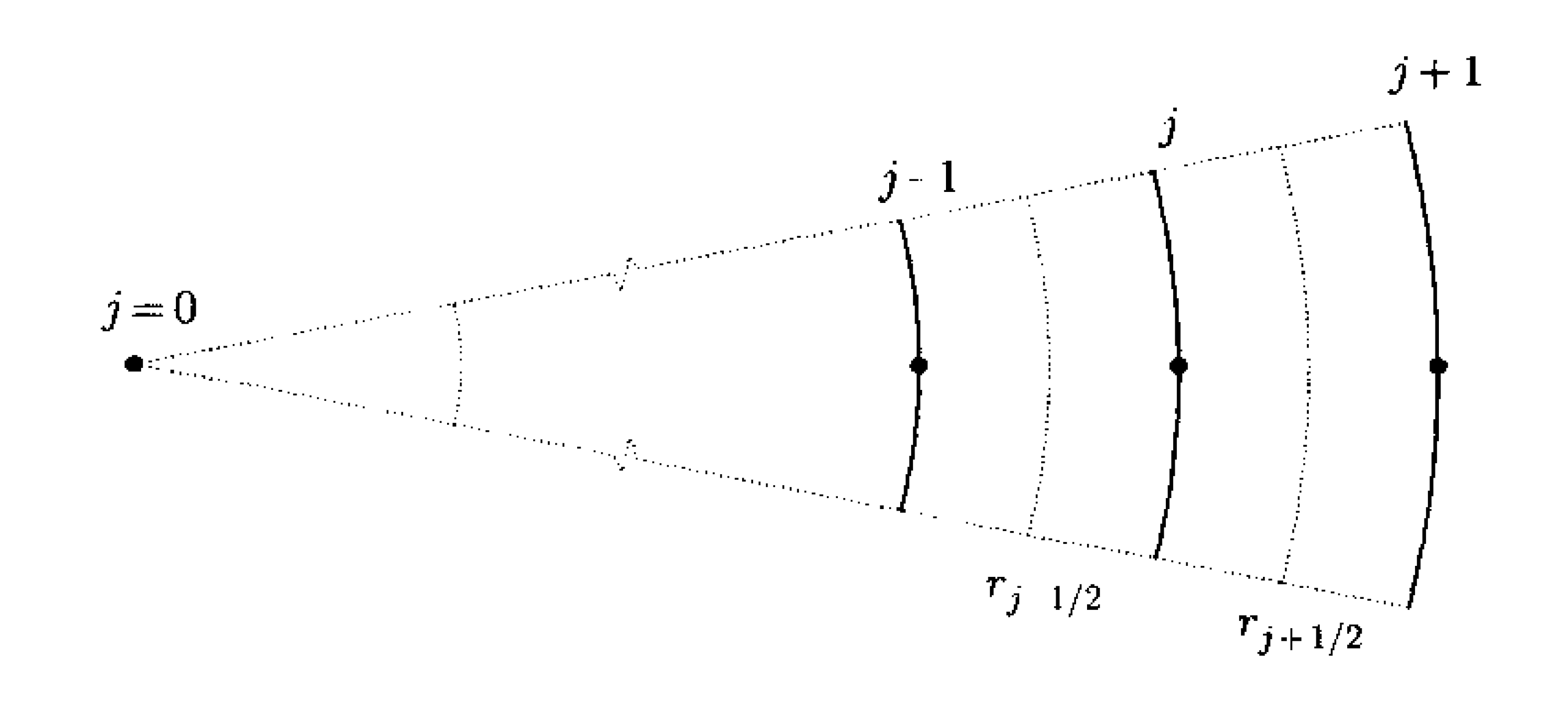


图 2-12　极坐标

因此在步长为 Δr 的一致网格上，对 (2.156) 的右端直接作差分得到

$$\begin{aligned}\frac{\partial U_j}{\partial t} &\approx \frac{\alpha+1}{r_{j+1/2}^{\alpha+1}-r_{j-1/2}^{\alpha+1}}\delta_r\left(r_j^{\alpha}\frac{\delta_r U_j}{\Delta r}\right)\\ &= \frac{(\alpha+1)\left[r_{j+1/2}^{\alpha}U_{j+1}-\left(r_{j+1/2}^{\alpha}+r_{j-1/2}^{\alpha}\right)U_j+r_{j-1/2}^{\alpha}U_{j-1}\right]}{\left[r_{j+1/2}^{\alpha}+r_{j+1/2}^{\alpha-1}r_{j-1/2}+\cdots+r_{j-1/2}^{\alpha}\right](\Delta r)^2},\end{aligned} \tag{2.162a}$$

其中 $j=1,2,\ldots$.

在极点处只有一个表面 (当 $\alpha=1$ 时是一个半径为 $r_{1/2}=\frac{1}{2}\Delta r$ 的圆柱面，当 $\alpha=2$ 时是一个半径为 $r_{1/2}$ 的球面), 则立即有

$$\frac{\partial U_0}{\partial t}\approx\frac{\alpha+1}{r_{1/2}^{\alpha+1}}r_{1/2}^{\alpha}\frac{U_1-U_0}{\Delta r}=2(\alpha+1)\frac{U_1-U_0}{(\Delta r)^2}, \tag{2.162b}$$

这与 (2.161) 是一致的. 同时注意在柱对称 ($\alpha=1$) 的情况下，(2.162a) 和不论由 (2.156) 或 (2.157) 得到的差分格式均是一致的；但是在球对称的情况下，因为 $r_{j+1/2}^2+r_{j+1/2}r_{j-1/2}+r_{j-1/2}^2$ 与 $3r_j^2$ 并不相等，因此它们是不一致的.

考虑最大值原理成立的条件，从 (2.162a) 和 (2.162b) 入手是最简单的. 计算 θ- 方法中 U_j^n 的系数，容易知道最差的情形发生在极点处并且得到条件

$$2(\alpha+1)(1-\theta)\Delta t\leqslant(\Delta r)^2. \tag{2.162c}$$

随着空间维数的增加这个条件的限制也越来越苛刻，不过其变化趋势与我们将在第 3 章讨论的内容是一致的.

2.17　非线性问题

在 2.15 节中考虑的一般线性方程的物理性质依赖于 x 和 t. 这些性质依赖于未知函

数 $u(x,t)$ 的情形也是很常见的. 这自然引起了对非线性问题的研究.

我们只考虑一个例子，即方程

$$u_t = b(u)u_{xx} \tag{2.163}$$

其中系数 $b(u)$ 仅依赖于解 u, 并且假定其对所有的 u 值都是严格正的. 这种简化实际上仅仅是为了简化记号；处理 b 同时是 x, t 和 u 的函数的情况并不比这种情况困难多少.

显式格式几乎没有变化；使用和前面相同的符号，得格式

$$U_j^{n+1} = U_j^n + \mu' b(U_j^n)\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right). \tag{2.164}$$

实际的计算不比线性情况更困难，唯一额外增加的工作量是函数值 $b(U_j^n)$ 的计算. 截断误差的形式也是完全一样的，并且 $\{U_j^n\}$ 满足最大值原理的条件也没有改变. 然而，全局误差 e_j^n 性态的分析却更困难，因为随着 n 的增长它是非线性变化的.

记精确解 $u(x_j, t_n)$ 为 u_j^n, 我们已知 U_j^n 和 u_j^n 分别满足方程

$$U_j^{n+1} = U_j^n + \mu' b(U_j^n)(U_{j+1}^n - 2U_j^n + U_{j-1}^n), \tag{2.165}$$

$$u_j^{n+1} = u_j^n + \mu' b(u_j^n)(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t T_j^n, \tag{2.166}$$

其中 T_j^n 是截断误差. 但是简单的把这两个方程相减不会得到 e_j^n 的关系式，因为两个方程的系数 $b(\cdot)$ 是不同的. 不过，我们可以先将 $b(u_j^n)$ 写为

$$b(u_j^n) = b(U_j^n) + (u_j^n - U_j^n)\frac{\partial b}{\partial u}(\eta) \tag{2.167}$$

$$= b(U_j^n) - e_j^n q_j^n, \tag{2.168}$$

其中

$$q_j^n = \frac{\partial b}{\partial u}(\eta), \tag{2.169}$$

且 η 是 U_j^n 和 u_j^n 之间的某个值.

这样，从 (2.165) 中减去 (2.166), 便得

$$\begin{aligned} e_j^{n+1} = {} & e_j^n + \mu' b(U_j^n)(e_{j+1}^n - 2e_j^n + e_{j-1}^n) \\ & + \mu' e_j^n q_j^n (u_{j+1}^n - 2u_j^n + u_{j-1}^n) - \Delta t\, T_j^n. \end{aligned} \tag{2.170}$$

上式右端头两项中的 e_{j-1}^n, e_j^n 和 e_{j+1}^n 的系数，当网格尺度满足条件

$$\Delta t\left[\max b(U_j^n)\right] \leqslant \frac{1}{2}(\Delta x)^2 \tag{2.171}$$

时，都是非负的. 这就是新的稳定性条件，同时也是保证逼近解满足最大值原理的条件；一般在每个时间步都需要检验该条件并调整时间步. 不过，假定存在一个常值步长 Δt, 它对任何 j 和 n 都满足 (2.171); 并且假定有边界

$$\left|u_{j+1}^n - 2u_j^n + u_{j-1}^n\right| \leqslant M_{xx}(\Delta x)^2, \ \left|q_j^n\right| \leqslant K, \tag{2.172}$$

则我们有 (沿用先前的符号)

$$E^{n+1} \leqslant [1 + KM_{xx}\Delta t]\, E^n + \Delta t\, T. \tag{2.173}$$

更进一步，由于

$$(1 + KM_{xx}\Delta t)^n \leqslant \mathrm{e}^{KM_{xx}n\Delta t} \leqslant \mathrm{e}^{KM_{xx}t_F}, \tag{2.174}$$

这样我们就得到一个与 T 有关的全局误差界.

然而，虽然稳定性条件 (2.171) 并不比线性问题的稳定性条件的约束性强多少，但是误差界却要差许多，除非可以先验地估计出 $|\partial b/\partial u|$ 和 $|u_{xx}|$ 的界都非常小. 再说，例 (2.163) 的形式相当特殊；同样常见的是 $u_t = (b(u)u_x)_x$，它多出的额外项 $(\partial b/\partial u)(u_x)^2$ 会对问题及其分析产生巨大影响.

在实际应用中偏微分方程的数值方法主要用于求解非线性问题，而这时其他方法一般是无效的. 如上所述，在实际的非线性问题中使用显式格式不会遇到多少困难；事实上，甚至隐式格式的使用也不是很困难的，只是需要迭代求解一个非线性方程组，这时前一时间层得到的数值解给出了好的迭代初值. 然而，这些格式的收敛性和稳定性分析比线性情况要困难的多.

文献注记与推荐读物

作为本书中关于初值问题的全部材料可参见 Richtmyer 和 Morton (1967) 的经典教材. 另一本 Collatz(1966) 的经典教材涵盖了各种偏微分方程问题. 有关本章所讨论问题的许多差分逼近格式可以在这两本书中找到.

要广泛了解扩散问题和它们的应用背景，读者可以参考 Crank(1975) 的书，其中可以找到对 Crank-Nicolson 格式应用的更多讨论.

关于重要的 Thomas 算法，我们所知的最早的参考文献是 1949 年来自纽约哥伦比亚大学的一个报告；但是报告中采用的方法相当于没有选主元的直接高斯消去法，在当时这种算法无疑已被许多研究者所知. 对用高斯消去法处理带状矩阵的更一般的讨论，读者可以参考数值分析的标准教材，在本书末尾的参考文献中罗列了若干，或者矩阵计算方面更专门的书籍，如 Golub 和 Van Loan(1996) 所著的书籍.

在 Ames(1992) 的书中可以找到对非线性问题更加完整的讨论，其参考文献 Ames(1965) 和 Ames(1972) 的书中例举了许多以非线性抛物型方程为模型的物理问题的例子.

习 题

2.1 (i) 函数 $u^0(x)$ 在 $[0,1]$ 上定义为,

$$u^0(x)=\begin{cases}2x, & \text{当 } 0\leqslant x\leqslant \frac{1}{2} \text{ 时},\\ 2-2x, & \text{当 } \frac{1}{2}\leqslant x\leqslant 1\text{时}.\end{cases}$$

证明

$$u^0(x)=\sum_{m=1}^{\infty}a_m\sin m\pi x,$$

其中 $a_m=(8/m^2\pi^2)\sin\frac{1}{2}m\pi$.

(ii) 证明

$$\int_{2p}^{2p+2}\frac{1}{x^2}\mathrm{d}x>\frac{2}{(2p+1)^2},$$

并且进而有

$$\sum_{p=p_0}^{\infty}\frac{1}{(2p+1)^2}<\frac{1}{4p_0}.$$

(iii) 用第 (i) 部分中的正弦级数的前 $m=405$ 项之和逼近 $u^0(x)$, 证明在区间 $[0,1]$ 上, 其误差小于 0.001.

2.2 (i) 证明对每个正值 $\mu=\Delta t/(\Delta x)^2$, 存在常数 $C(\mu)$, 使得对所有的正值 k 和 Δx 有

$$\left|1-4\mu\sin^2\frac{1}{2}k\Delta x-\mathrm{e}^{-k^2\Delta t}\right|\leqslant C(\mu)k^4(\Delta t)^2.$$

验证当 $\mu=\frac{1}{4}$ 时有 $C=\frac{1}{2}$ 使上面不等式成立.

(ii) 在区域 $0\leqslant x\leqslant 1$, $t\geqslant 0$ 上用显式中心差分格式逼近方程 $u_t=u_{xx}$. 边界条件为 $u(0,t)=u(1,t)=0$, 初始条件和习题 2.1 中一样为 $u(x,0)=u^0(x)$. 取 $\epsilon=0.01$, 证明习题 2.1 的正弦级数满足

$$\sum_{m=2p_0+1}^{\infty}|a_m|\leqslant\frac{\epsilon}{4},\qquad \text{如果 } p_0=82,$$

并且有

$$\sum_{m=1}^{2p_0-1}|a_m|m^4\leqslant 8p_0(2p_0+1)(2p_0-1)/3\pi^2.$$

而且证明当 $\mu=\frac{1}{4}$ 时若有 $\Delta t\leqslant 1.7\times10^{-10}$, 则在 $0\leqslant t\leqslant 1$ 的时间范围内数值解的误差小于 0.01.

(iii) 取 $\mu=\frac{1}{4}$ 和 $\Delta t=0.0025$, 用数值计算验证数值解在这个时间范围内的误差小于 0.01. [注意这个模型问题中误差最大处始终出现在第一个时间步.]

2.3 假设网格点 x_j 满足

$$0 = x_0 < x_1 < x_2 < \cdots < x_{J-1} < x_J = 1,$$

但除此之外没有其他限制. 在时间范围 $0 \leqslant t \leqslant t_F$ 内, 用差分格式

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{2}{\Delta x_{j-1} + \Delta x_j}\left(\frac{U_{j+1}^n - U_j^n}{\Delta x_j} - \frac{U_j^n - U_{j-1}^n}{\Delta x_{j-1}}\right)$$

逼近方程 $u_t = u_{xx}$, 其中 $\Delta x_j = x_{j+1} - x_j$. 证明逼近的截断误差主项是

$$\begin{aligned} T_j^n &= \frac{1}{2}\Delta t\, u_{tt} - \frac{1}{3}(\Delta x_j - \Delta x_{j-1})u_{xxx} \\ &\quad - \frac{1}{12}\left[(\Delta x_j)^2 + (\Delta x_{j-1})^2 - \Delta x_j \Delta x_{j-1}\right] u_{xxxx}. \end{aligned}$$

现在假设已经指定边界条件 $u(0,t)$, $u(1,t)$ 和初始条件 $u(x,0)$. 记 $\Delta x = \max \Delta x_j$. 且假设网格是足够光滑的, 即对 $j = 1, 2, \ldots, J-1$ 有 $|\Delta x_j - \Delta x_{j-1}| \leqslant \alpha(\Delta x)^2$, 其中 α 是常数. 使用通常的符号, 求证若稳定性条件

$$\Delta t \leqslant \frac{1}{2}\Delta x_{j-1}\Delta x_j,\ j = 1, 2, \ldots, J-1$$

成立, 则有

$$\left|U_j^n - u(x_j, t_n)\right| \leqslant \left(\frac{1}{2}\Delta t M_{tt} + (\Delta x)^2\left\{\frac{1}{3}\alpha M_{xxx} + \frac{1}{12}[1 + \alpha\Delta x]M_{xxxx}\right\}\right) t_F.$$

2.4 已知 a_j, b_j, c_j 满足

$$a_j > 0,\ c_j > 0,\ b_j > a_j + c_j,\ j = 1, 2, \ldots, J-1,$$

和

$$e_j = \frac{c_j}{b_j - a_j e_{j-1}},\ j = 1, 2, \ldots, J-1,$$

其中 $e_0 = 0$. 归纳证明 $0 < e_j < 1,\ j = 1, 2, \ldots, J-1$. 进一步, 证明若有条件

$$b_j > 0,\ b_j \geqslant |a_j| + |c_j|,\ j = 1, 2, \ldots, J-1,$$

则从 $|e_0| \leqslant 1$ 可以推出 $|e_j| \leqslant 1,\ j = 1, 2, \ldots, J-1$.

2.5 方程 $u_t = u_{xx}$ 有边界条件 $u(1,t) = 0$ 对所有的 $t \geqslant 0$, 和

$$\frac{\partial u}{\partial x} = \alpha(t)u + g(t) \quad \text{在 } x = 0 \text{ 处，对任何 } t \geqslant 0,$$

其中 $\alpha(t) \geqslant 0$. 详细说明在用 θ- 方法求解方程过程中如何使用 Thomas 算法. 特别地, 推导出取代方程 (2.71) 的初始条件.

2.6 (i) 分别考虑实根和复根的情况, 证明实系数二次方程 $z^2 + bz + c = 0$ 的两个根在单位圆周内或者其上的充分必要条件是 $|c| \leqslant 1$ 和 $|b| \leqslant 1 + c$.

(ii) 证明对任意 μ 值，格式

$$U_j^{n+1} - U_j^{n-1} = \frac{1}{3}\mu\left\{\delta_x^2 U_j^{n+1} + \delta_x^2 U_j^n + \delta_x^2 U_j^{n-1}\right\}$$

都是稳定的.

(iii) 证明无论 μ 取何值，格式

$$U_j^{n+1} - U_j^{n-1} = \frac{1}{6}\mu\left\{\delta_x^2 U_j^{n+1} + 4\delta_x^2 U_j^n + \delta_x^2 U_j^{n-1}\right\}$$

都是不稳定的.

2.7 在区域 $0 < x < 1$, $t > 0$ 上有微分方程

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right),$$

边界条件为在 $x = 0$ 和 $x = 1$ 处给定 u 的值. 求逼近方程的显式格式

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{\left\{(U_{j+1}^n - U_j^n)p_{j+1/2} - (U_j^n - U_{j-1}^n)p_{j-1/2}\right\}}{(\Delta x)^2}$$

截断误差的主项. 设条件 $0 < p(x)\Delta t \leqslant \frac{1}{2}(\Delta x)^2$ 成立，推出用 u 和 p 的导数界表示的数值格式全局误差的上界.

2.8 对上一习题中的问题使用 θ-方法, 证明如果 $p(x) > 0$ 则 Thomas 算法稳定性的条件即成立. 同时证明，若对所有的 x, 条件 $2\Delta t(1-\theta)p(x) \leqslant (\Delta x)^2$ 成立，则最大值原理成立.

2.9 用 θ- 方法逼近方程 $u_t = u_{xx}$, 选取

$$\theta = \frac{1}{2} + \frac{(\Delta x)^2}{12\Delta t}.$$

证明得到的格式是无条件稳定的，其截断误差是 $O((\Delta t)^2 + (\Delta x)^2)$ 阶的，并且它对所有的随时间步振荡的傅里叶波型比 Crank-Nicolson 格式衰减的更快. 然而，为保证最大值原理成立，证明网格比 $\Delta t/(\Delta x)^2$ 必须在 $[\frac{1}{6}, \frac{7}{6}]$ 范围内.

2.10 在 x 方向使用非一致网格求解方程 $u_t = u_{xx}$, 网格如下

$$x_j = \frac{j^2}{J^2}, \quad j = 0, 1, 2, \ldots, J.$$

通过变量替换 $x = s^2$ 将方程形式变为

$$u_t = \frac{1}{2s}\frac{\partial}{\partial s}\left(\frac{1}{2s}\frac{\partial u}{\partial s}\right);$$

使用 $\Delta s = 1/J$ 的一致网格，并且使用习题 2.7 中的格式，格式仅在右端多出额外因子 $1/2s_j$. 证明其截断误差的主项是

$$T_j^n = \frac{1}{2}\Delta t\, u_{tt} - \frac{1}{24}(\Delta s)^2\frac{1}{2s}\left[(\frac{1}{2s}u_{sss})_s + (\frac{1}{2s}u_s)_{sss}\right],$$

并且由变量替换可将其变换为

$$T_j^n = \frac{1}{2}\Delta t\, u_{tt} - (\Delta s)^2\left(\frac{2}{3}u_{xxx} + \frac{1}{3}x\, u_{xxxx}\right).$$

且比较它和习题 2.3 中得到的截断误差主项.

2.11 用 Crank-Nicolson 格式求解方程 $u_t = u_{xx}$. 边界条件为 $U_0^n = U_J^n = 0,\ n \geqslant 0$; 初始条件为对某个固定 k 有 $U_k^0 = 1$, 其中 $0 < k < J$, 而对所有的 $j \neq k$ 有 $U_j^0 = 0$. 记 $w_j = U_j^1$, 验证 w_j 满足递归关系

$$-\frac{1}{2}\mu w_{j-1} + (1+\mu)w_j - \frac{1}{2}\mu w_{j+1} = q_j, \quad w_0 = w_J = 0,$$

其中 $q_k = 1-\mu$, $q_{k+1} = q_{k-1} = \frac{1}{2}\mu$, 而其他的 $q_j = 0$. 假设网格足够细, 并且 x_k 和边界足够远使得 k 和 $J-k$ 都足够大. 解释为什么

$$w_j = Ap^{|j-k|}, \qquad j \neq k,$$
$$w_k = A + B,$$

是 w_j 的一个好的近似, 其中 $p = (1+\mu-\sqrt{(1+2\mu)})/\mu$. 写出并求解关于常数 A 和 B 两个方程, 证明 $w_k = 2/\sqrt{(1+2\mu)} - 1$. 进而证明 (i) 对所有 $\mu > 0$ 有 $w_k < 1$; (ii) $w_k > 0$ 当且仅当 $\mu < \frac{3}{2}$; 和 (iii) $w_k \geqslant w_{k+1}$ 当且仅当 $\mu \leqslant (7-\sqrt{17})/4$.

2.12 证明 (Hermitian) 差分格式

$$\begin{aligned}\left(1+\frac{1}{12}\delta_x^2\right)\left(U^{n+1}-U^n\right) &= \frac{1}{2}\mu\,\delta_x^2\left(U^{n+1}+U^n\right)\\ &\quad +\frac{1}{2}\Delta t\left[f^{n+1} + (1+\frac{1}{6}\delta_x^2)f^n\right]\end{aligned}$$

逼近 $u_t = u_{xx} + f$ 的截断误差是 $O\left((\Delta t)^2\right)$ 阶的, 方程中 f 为一个给定的函数, 并且 $\mu = \Delta t/(\Delta x)^2$ 取固定值.

第 3 章　二维和三维抛物型方程

3.1　盒形区域上的显式方法

一维模型问题在二维的自然推广就是方程

$$\begin{aligned} u_t &= b\nabla^2 u \quad (b>0) \\ &= b[u_{xx}+u_{yy}], \end{aligned} \tag{3.1}$$

其中 b 是正常数. 考虑 $(x,\ y)$ 平面上的矩形区域

$$0<x<X, \quad 0<y<Y,$$

且假定取狄利克雷边界条件，即对所有的 $t>0$, $u(x,y,t)$ 在矩形区域的边界上取给定的值. 当然，还要给定初始条件，即在矩形区域上 $u(x,y,0)$ 是给定的. 在该区域上引入均匀分布的矩形网格点，其 x 方向的步长为 Δx, y 方向的步长为 Δy, 其中

$$\Delta x = \frac{X}{J_x}, \quad \Delta y = \frac{Y}{J_y}, \qquad J_x,\ J_y \in \mathbb{Z}.$$

于是近似解 (approximate solution) 可以记为

$$U^n_{r,s} \approx u(x_r, y_s, t_n), \quad r=0,1,\ldots,J_x, \quad s=0,1,\ldots,J_y.$$

最简单的二维显式差分格式是一维显式差分格式的自然推广，由下式给出

$$\frac{U^{n+1}-U^n}{\Delta t} = b\left[\frac{\delta_x^2 U^n}{(\Delta x)^2} + \frac{\delta_y^2 U^n}{(\Delta y)^2}\right]. \tag{3.2}$$

这里省略了下标 $(r,\ s)$, 且使用了 (2.28) 式中关于 x 和 y 方向二阶中心差分的符号. 由于在新的时间层上有唯一的未知量 $U^{n+1}_{r,s}$, 因此该格式是显式的，这个未知量可由前一时间层上五个邻近点的值

$$U^n_{r,s},\ U^n_{r+1,s},\ U^n_{r-1,s},\ U^n_{r,s+1} \text{ 和 } U^n_{r,s-1} \tag{3.3}$$

计算得到. 该格式一维时的大多数分析结果都可以简单地推广到二维情形，有关的细节留作练习. 二维问题的截断误差为

$$T(x,t) = \frac{1}{2}\Delta t\, u_{tt} - \frac{1}{12}b\left[(\Delta x)^2 u_{xxxx} + (\Delta y)^2 u_{yyyy}\right] + \cdots, \tag{3.4}$$

由此知截断误差的界可以用 u 的各阶导数的界给出，用前一章的符号，这些相应的导数

界记为 M_{tt}, M_{xxxx} 和 M_{yyyy}. 类似地可以证明收敛性，推出误差估计式

$$E^n \leqslant \left[\frac{1}{2}\Delta t\, M_{tt} + \frac{1}{12}b\left((\Delta x)^2 M_{xxxx} + (\Delta y)^2 M_{yyyy}\right)\right] t_F, \tag{3.5}$$

当网格尺度满足条件

$$\mu_x + \mu_y \leqslant \frac{1}{2}, \tag{3.6}$$

时成立，其中

$$\mu_x = \frac{b\Delta t}{(\Delta x)^2},\ \ \mu_y = \frac{b\Delta t}{(\Delta y)^2}. \tag{3.7}$$

假设 b 为常数且忽略边界条件的影响，则同样可以用傅里叶级数来分析该格式的稳定性. 至于为什么可以忽略边界条件的影响，第 5 章中将会给出解释. 构造如下形式的差分方程的解

$$U^n \sim (\lambda)^n \exp(\mathrm{i}\,[k_x x + k_y y]), \tag{3.8}$$

由此可得增长因子 λ 的表达式为

$$\lambda \equiv \lambda(\mathbf{k}) = 1 - 4\left[\mu_x \sin^2 \frac{1}{2}k_x\Delta x + \mu_y \sin^2 \frac{1}{2}k_y\Delta y\right], \tag{3.9}$$

其中 $\mathbf{k} = (k_x,\ k_y)$. 显然，恰如一维情形，稳定性的条件是

$$\mu_x + \mu_y \leqslant \frac{1}{2}. \tag{3.10}$$

该条件在证明 $|\lambda(\mathbf{k})| \leqslant 1$ 时的充分性与一维情形完全类似；而以下事实则证明了其必要性： $\mathbf{k}$ 的各分量可以独立选取，而可以取到的最坏波型的条件是 $k_x\Delta x = k_y\Delta y = \pi$.

用 (3.2) 式计算近似解显然与一维情形一样简单. 但稳定性条件的约束性却更大，而且当 b 是变量时条件 (3.10) 必须逐点满足，以便使 b 的任何局部峰值减小可用的时间步. 因此这种显式格式一般并不实用，我们必须引入某种隐性格式以避免，或放松稳定性的约束. 以上的讨论很容易推广到三维情形，事实上三维问题的隐式格式尤为重要.

3.2　二维 ADI 方法 (交替方向迭代法)

我们来考虑 θ-方法 的推广，它作为一维问题研究方法的自然推广之一. 特别地，Crank-Nicolson 方法变成了

$$(1 - \frac{1}{2}\mu_x\delta_x^2 - \frac{1}{2}\mu_y\delta_y^2)U^{n+1} = (1 + \frac{1}{2}\mu_x\delta_x^2 + \frac{1}{2}\mu_y\delta_y^2)U^n. \tag{3.11}$$

这类方法在一维情形几乎不需要增加额外的计算量就可以取消对稳定性的约束，因而具有很大的优越性. 对二维和三维问题情况并非如此，尽管对任何时间步长这类方法仍然

是稳定的，但因此而增加的工作量却也十分可观. 这时必须求解包含 $(J_x-1)(J_y-1)$ 个未知量 $U_{r,s}^{n+1}$ 的线性方程组. 该方程组有规则的结构，每个方程最多包含 5 个未知量；方程组的系数矩阵有大量的零元素，但却没有三对角 (tridiagonal) 的形式；事实上不可能通过行和列的置换将其化为由非零元素组成的带宽很窄的矩阵. 当然这类方程组并非无法求解，第 6 章中讨论椭圆型方程时我们将会看到其求解的方法，然而这样做带来了太多额外的复杂性，这就促使我们去寻找求解二维抛物型方程的其他数值格式.

由于隐式格式在一维时会非常有效， 那么自然就会想到设计仅关于一维，而不是两维，为隐式的格式. 例如，考虑格式

$$(1-\frac{1}{2}\mu_x\delta_x^2)U^{n+1}=(1+\frac{1}{2}\mu_x\delta_x^2+\mu_y\delta_y^2)U^n. \tag{3.12}$$

如果我们考察该方程组中相应于某一特定网格点的行或 s 值的那些方程，就会发现这些方程不包含任何具有不同 s 值的未知量，因而形成了 J_x-1 阶的三对角方程组. 于是整个方程组就包含了 J_y-1 个三对角方程组，而每个三对角方程组都可以有效地用 2.9 节中的 Thomas 算法求解. 该格式的计算工作量大约是显式格式的 3 倍. 遗憾地是，尽管格式的稳定性得到了改善，但其约束依然存在. 仍设 b 为常数，则易证格式的增长因子为

$$\lambda(\mathbf{k})=\frac{1-2\mu_x\sin^2\frac{1}{2}k_x\Delta x-4\mu_y\sin^2\frac{1}{2}k_y\Delta y}{1+2\mu_x\sin^2\frac{1}{2}k_x\Delta x},$$

当 $\mu_y>\frac{1}{2}$ 时，格式不稳定. 正如所料，稳定性对于 μ_x 没有任何约束.

将两个这种沿一个方向为隐式的格式结合起来就可以得到成功的解决问题的方法. 第一个这类格式是由 Peaceman 和 Rachford 在 1955 年 [1] 提出并应用于油藏模拟. 我们先写出改进的 Crank-Nicolson 格式

$$(1-\frac{1}{2}\mu_x\delta_x^2)(1-\frac{1}{2}\mu_y\delta_y^2)U^{n+1}=(1+\frac{1}{2}\mu_x\delta_x^2)(1+\frac{1}{2}\mu_y\delta_y^2)U^n. \tag{3.13}$$

注意到微分算子的乘积可展开为

$$(1+\frac{1}{2}\mu_x\delta_x^2)(1+\frac{1}{2}\mu_y\delta_y^2)=(1+\frac{1}{2}\mu_x\delta_x^2+\frac{1}{2}\mu_y\delta_y^2+\frac{1}{4}\mu_x\mu_y\delta_x^2\delta_y^2), \tag{3.14}$$

可以看出 (3.13) 式与 Crank-Nicolson 格式并不完全一样，不过其中多出来的几项与截断误差中的某些项同阶 (见下面的 (3.19) 式). 引进中间层变量 $U^{n+1/2}$, 则 (3.13) 式可以写成以下等价形式

$$(1-\frac{1}{2}\mu_x\delta_x^2)U^{n+1/2} = (1+\frac{1}{2}\mu_y\delta_y^2)U^n, \tag{3.15a}$$

$$(1-\frac{1}{2}\mu_y\delta_y^2)U^{n+1} = (1+\frac{1}{2}\mu_x\delta_x^2)U^{n+1/2}, \tag{3.15b}$$

[1] Peaceman, D.W. and Rachford, H.H. Jr (1955), The numerical solution of parabolic and elliptic differential equations, *J. Soc. Indust. Appl. Math.* **3**, 28-41.

要证明其等价性只需在 (3.15a) 式上作用 $(1+\frac{1}{2}\mu_x\delta_x^2)$，在 (3.15b) 式上作用 $(1-\frac{1}{2}\mu_x\delta_x^2)$ 即可.

(3.15a) 式的右端项由前一时间步得到；算出 $U^{n+1/2}$ 后，则 (3.15b) 式的右端项也得到了. 与单变量的隐式格式 (3.12) 完全一样，这两个方程组都是由若干三对角方程组构成的. 每个时间步的全部工作量包括先求解 J_y-1 个 J_x-1 阶的三对角方程组 (比如对应于图 3-1 中标注叉号的一组点)；然后求解 J_x-1 个 J_y-1 阶的三对角方程组 (比如对应于图 3-1 中标注圆点的一组点). 整个算法过程比求解由 Crank-Nicolson 方法得到的 $(J_x-1)(J_y-1)$ 阶的完整方程组要快得多；每个网格点需要大约 10(加) + 8(乘) + 6(除) 次运算，差不多是显式格式 (3.2) 的 4(加) + 3(乘) 次运算的 3 倍. 在图 3-1 中标注 □ 的点处需要给出 $U^{n+1/2}$ 的边界条件，而在图 3-1 中标注 ⊙ 的点处需要给出 U^n 的边界条件.

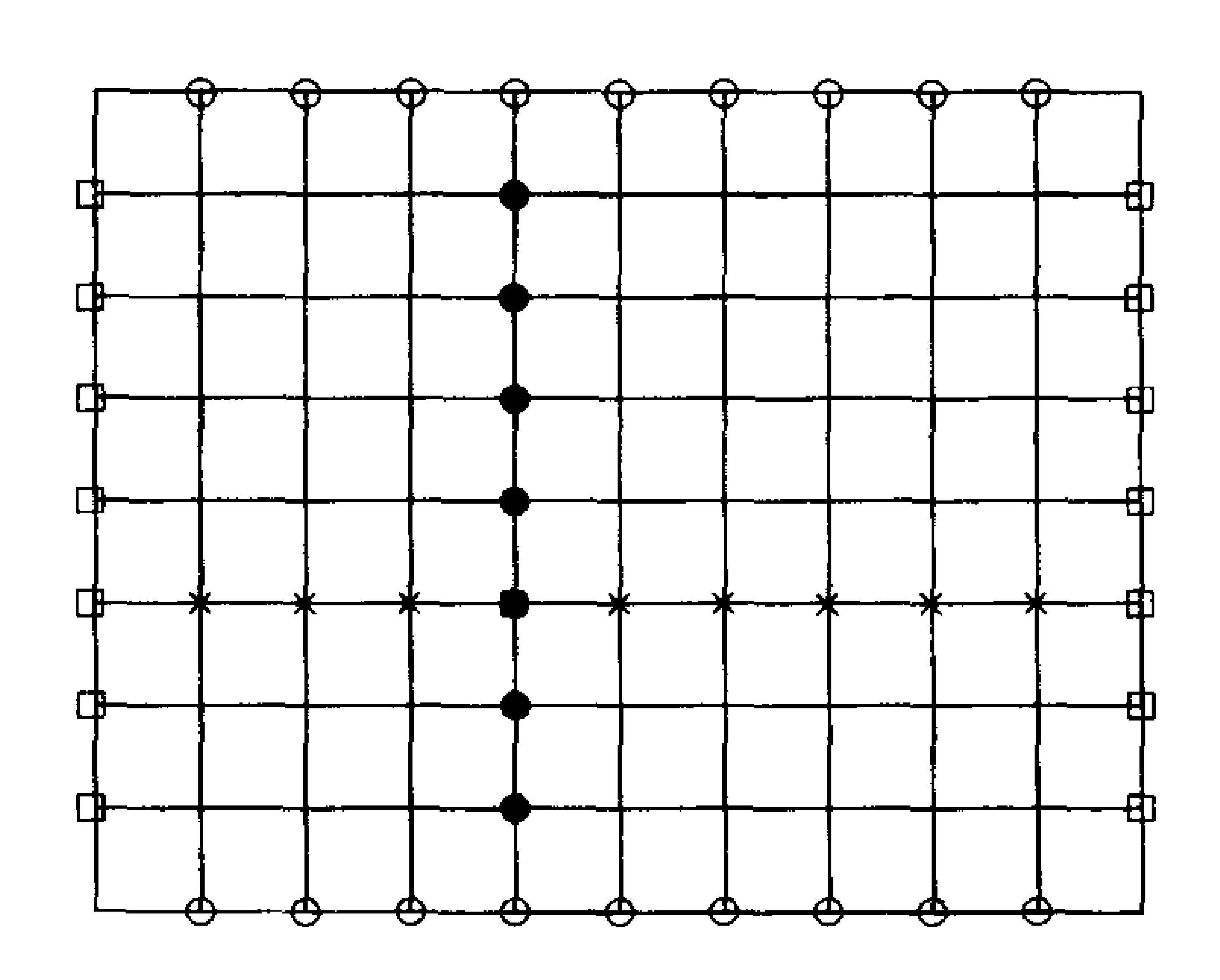


图 3-1　ADI 方法的边界点

当 b 为常数时，我们仍可以将傅里叶波型 (3.8) 式代入 (3.13) 式以分析其稳定性. 从两种等价形式都可以得到

$$\lambda(\mathbf{k})=\frac{(1-2\mu_x\sin^2\frac{1}{2}k_x\Delta x)(1-2\mu_y\sin^2\frac{1}{2}k_y\Delta y)}{(1+2\mu_x\sin^2\frac{1}{2}k_x\Delta x)(1+2\mu_y\sin^2\frac{1}{2}k_y\Delta y)}, \tag{3.16}$$

由此立即可知格式的无条件稳定性.

我们也可以将最大值原理应用于 (3.15) 式. 在前半个时间步上，每个方程可以写成

$$(1+\mu_x)U_{r,s}^{n+1/2}=(1-\mu_y)U_{r,s}^n+\frac{1}{2}\mu_y(U_{r,s-1}^n+U_{r,s+1}^n)+\frac{1}{2}\mu_x(U_{r+1,s}^{n+1/2}+U_{r-1,s}^{n+1/2}). \tag{3.17}$$

因此，当 $\mu_y\leqslant 1$ 时，$U_{r,s}^{n+1/2}$ 的值就可以表示为其邻近网点上 U^n 和 $U^{n+1/2}$ 取值的线性组合，组合的系数均为非负且和为 1. 对后半个时间步，显然有类似的结论. 于是由最大值原理便知，只要

$$\max\{\mu_x,\ \mu_y\}\leqslant 1, \tag{3.18}$$

就有数值解的取值必然落在其边界上的最大和最小值之间. 而 (3.18) 式则是一维 Crank-Nicolson 方法稳定性条件的自然推广.

计算截断误差最理想的做法是从非分裂形式的 (3.13) 式出发，把所有的项都移到右端后再除以 Δt, 就不难推出截断误差的主项为

$$\begin{aligned} T^{n+1/2} &\approx \tfrac{1}{24}(\Delta t)^2 u_{ttt} - \tfrac{1}{12}(\Delta x)^2 u_{xxxx} - \tfrac{1}{12}(\Delta y)^2 u_{yyyy} \\ &\quad -\tfrac{1}{8}(\Delta t)^2 u_{xxtt} - \tfrac{1}{8}(\Delta t)^2 u_{yytt} + \tfrac{1}{4}(\Delta t)^2 u_{xxyyt} \\ &= O((\Delta t)^2 + (\Delta x)^2 + (\Delta y)^2), \end{aligned} \tag{3.19}$$

前五项与二维情形的 Crank-Nicolson 格式相同 (一维情形见 (2.85) 式), 而最后一项源于乘积项 $\delta_x^2\delta_y^2(U^{n+1}-U^n)$.

作为例子，我们考虑在单位正方形区域 $0<x<1, 0<y<1$ 上的热传导方程

$$u_t = u_{xx} + u_{yy}, \tag{3.20}$$

并在单位正方形区域的边界上给定齐次狄利克雷边界条件 $u=0$, 初始条件为 $u(x,y,0)=f(x,y)$, 其中在如图 3-2 所示的形如字母 **M** 的区域内 $f(x,y)=1$, 而在单位正方形区域的其他地方 $f(x,y)=0$. 在 **M** 周围的窄条区域中，函数由 0 线性地增长到 1, 因此 $f(x,y)$ 是连续的，但其导函数不是连续的，事实上在窄条区域中导数很大，而在此之外导数为零.

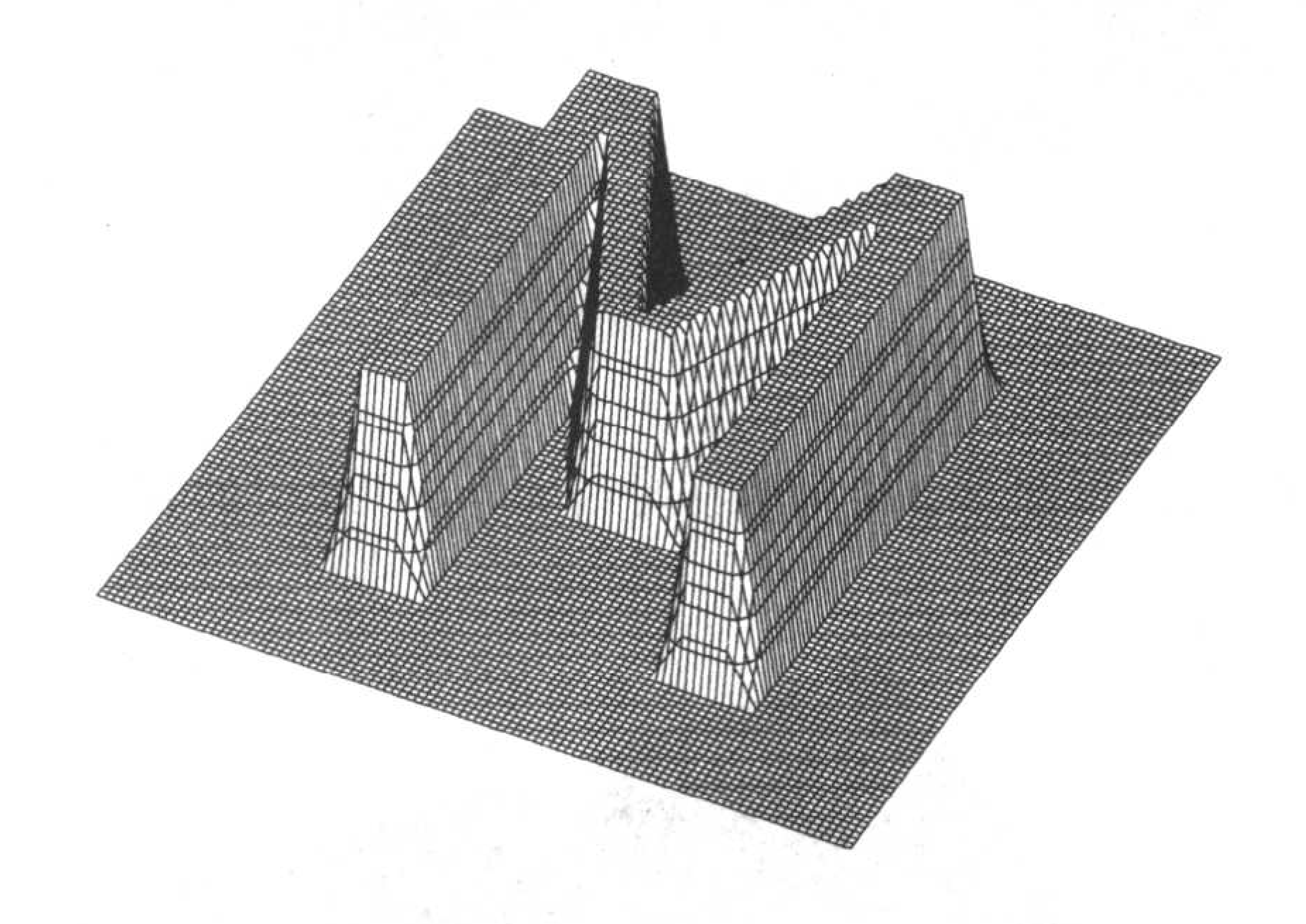


图 3-2 单位正方形区域上热流计算的初值

图 3-3 、图 3-4 和图 3-5 显示了显式格式数值计算的结果；由此可以看出初始函数是如何扩散到整个正方形区域上去的. 类似于第 2 章，原则上数值解的误差也可以通过与精确解的傅里叶级数展开相比较而得到. 傅里叶级数的系数可能极为复杂，因此我们采用的方法是用不同尺度的网格得到的计算结果来估计数值解的精度. 取 $\Delta x=\Delta y=1/100$, 1/200 和 1/400, 数值结果相互间吻合得很好，这使我们有信心相信图中显示的数值解的精度远高于该图像所表示的结果.



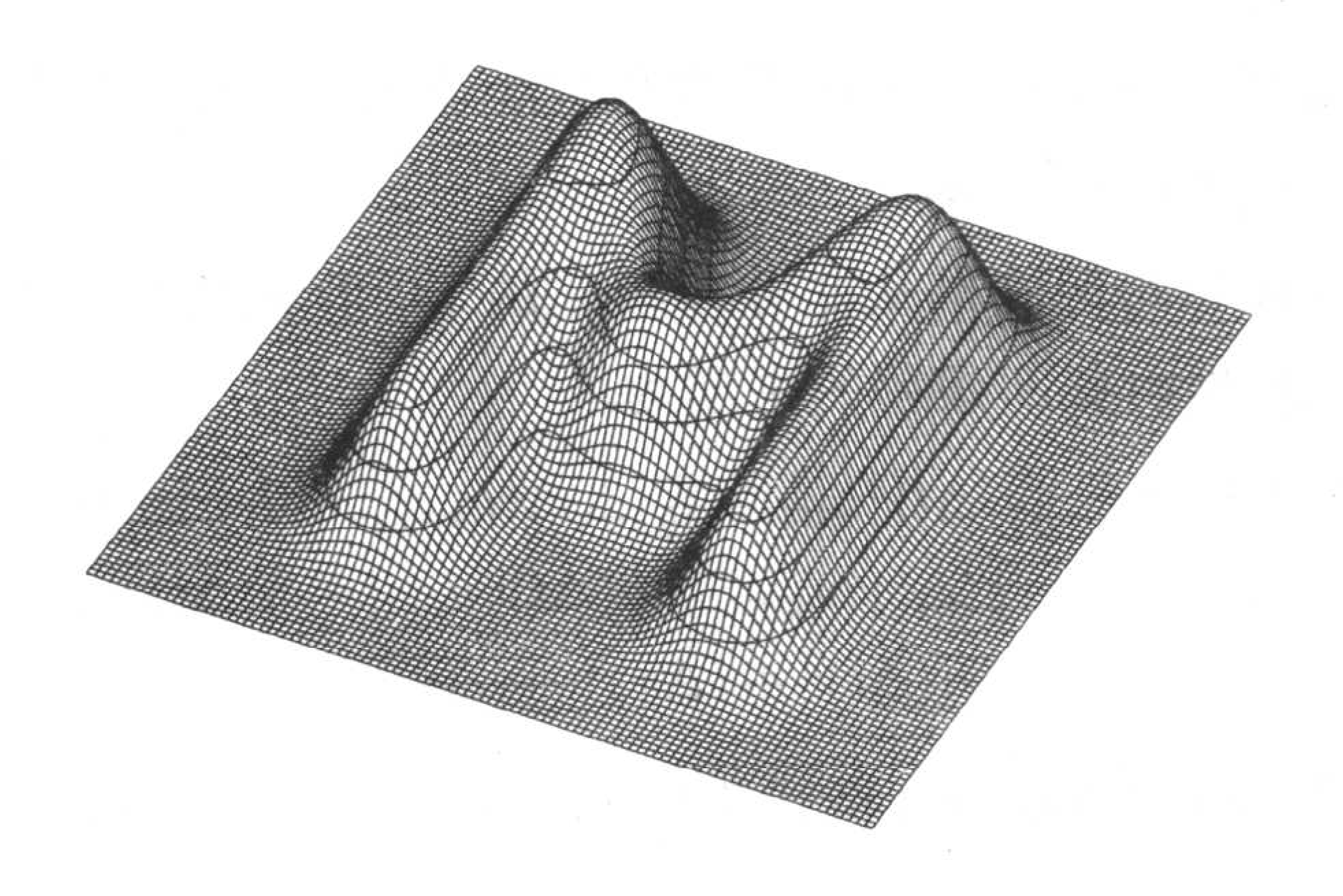


图 3-3 以图 3-2 为初值在时刻 $t = 0.001$ 时的数值解

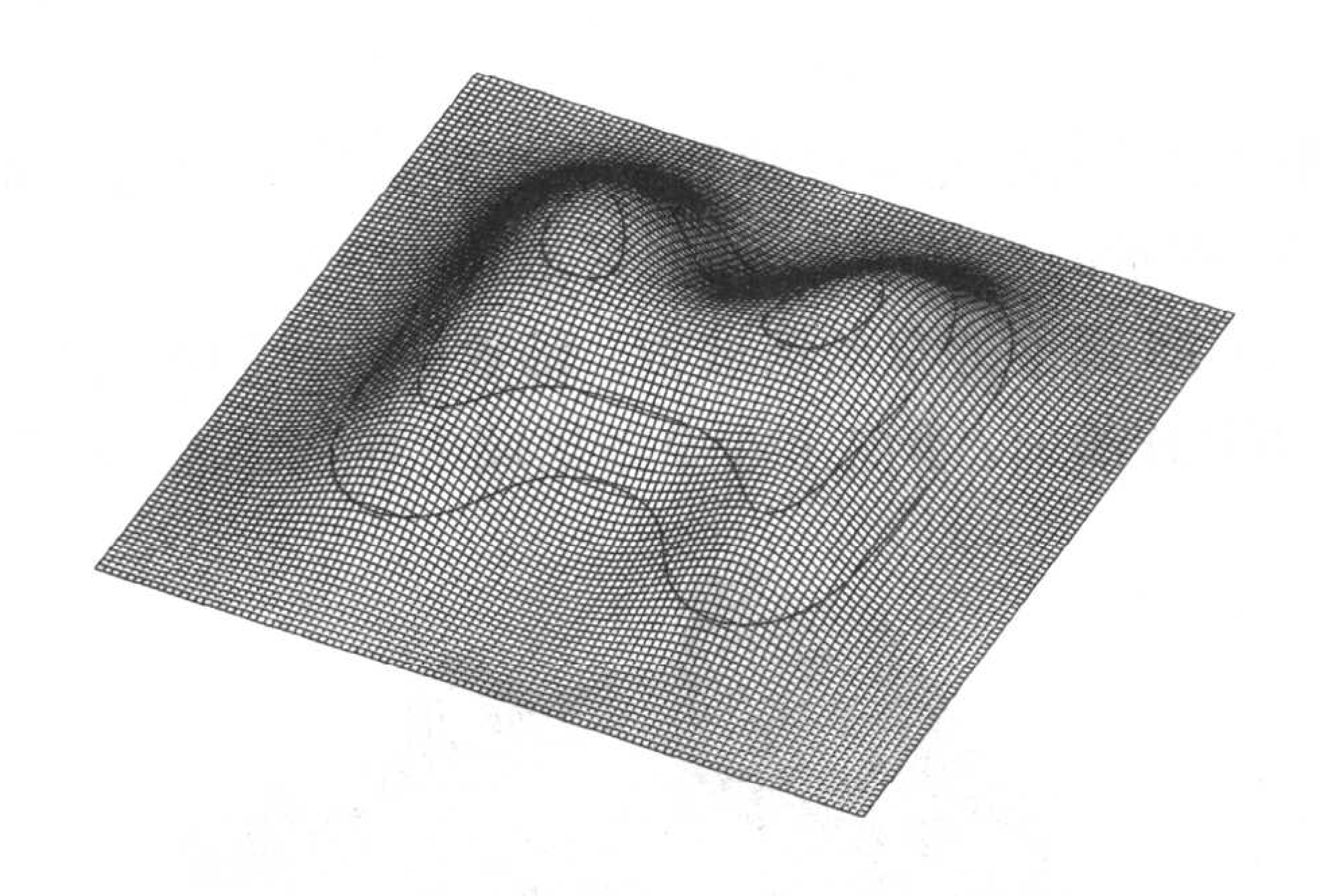


图 3-4 以图 3-2 为初值在时刻 $t = 0.004$ 时的数值解

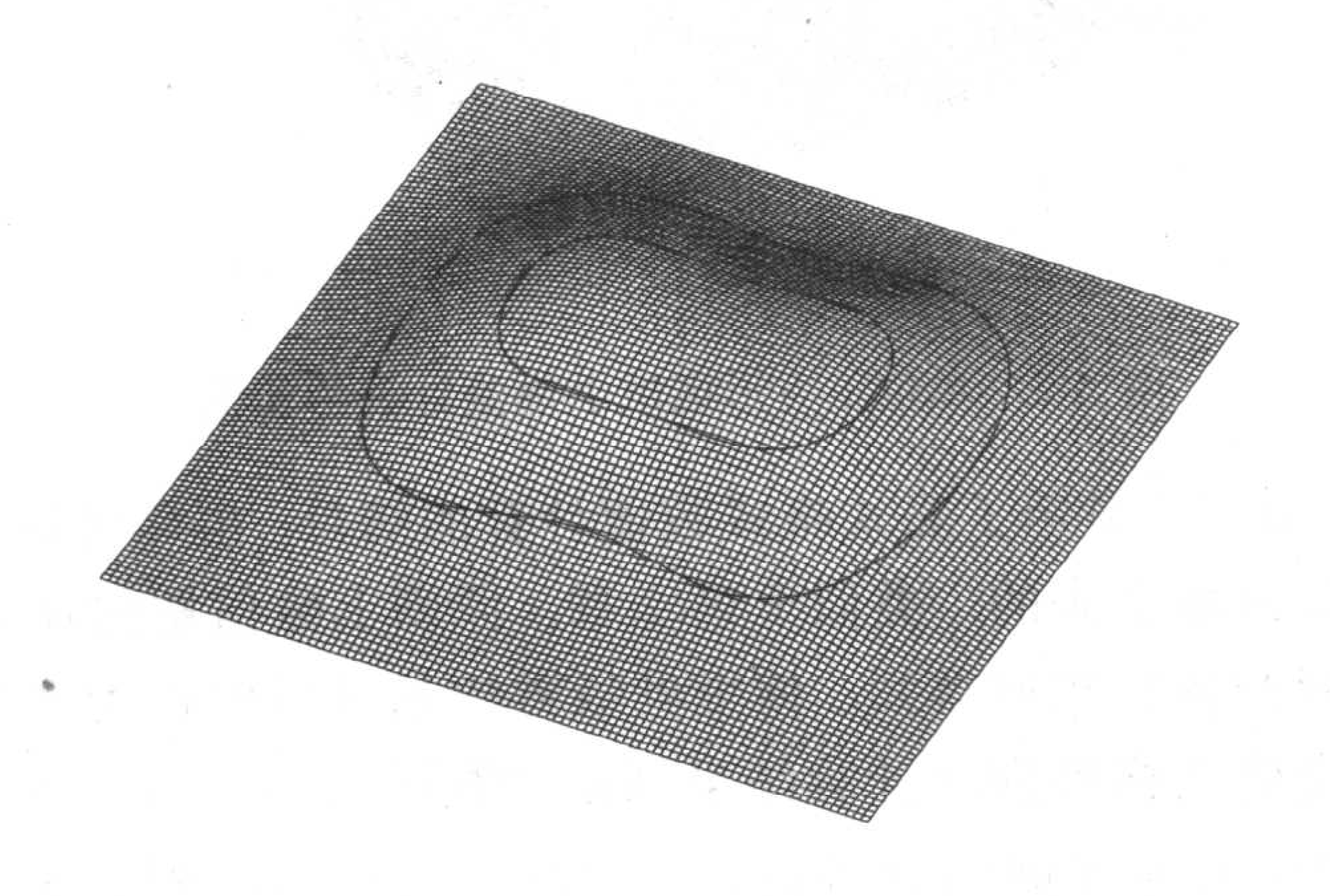


图 3-5 以图 3-2 为初值在时刻 $t = 0.01$ 时的数值解

3.3 三维 ADI 和 LOD 方法

有许多类似于 (3.15) 式的 ADI 方法 (alternating direction implicit method), 另外也有其他方法可以只求解三对角方程组而又具有隐式格式的优越性. 有时格式的差别仅在于中间量的计算, 这会影响边界条件的提法. 例如, D'yakonov[1] 提出把 (3.13) 式分成为

$$(1-\frac{1}{2}\mu_x\delta_x^2)U^{n+*} = (1+\frac{1}{2}\mu_x\delta_x^2)(1+\frac{1}{2}\mu_y\delta_y^2)U^n, \qquad (3.21a)$$

$$(1-\frac{1}{2}\mu_y\delta_y^2)U^{n+1} = U^{n+*}. \qquad (3.21b)$$

如记号所示, U^{n+*} 并非某中间时刻的相容的近似解, 因此必须特别注意其边界条件, 不过对 U^{n+*} 来说只需给出左右两侧 (图 3-1 中标注 □ 的点) 的边界条件, 而这些点上 U^{n+*} 的值可由 (3.21b) 式得到. 当考虑向三维问题推广时, 表示成不同形式的格式会带来一些好处.

一种常用的 ADI 方法是由 Douglas 和 Rachford[2] 提出的, 其二维形式为

$$(1-\mu_x\delta_x^2)U^{n+1*} = (1+\mu_y\delta_y^2)U^n, \qquad (3.22a)$$

$$(1-\mu_y\delta_y^2)U^{n+1} = U^{n+1*}-\mu_y\delta_y^2U^n; \qquad (3.22b)$$

消掉 U^{n+1*} 后即得

$$(1-\mu_x\delta_x^2)(1-\mu_y\delta_y^2)U^{n+1} = (1+\mu_x\mu_y\delta_x^2\delta_y^2)U^n. \qquad (3.23)$$

(3.15) 式是受 Crank-Nicolson 格式的启发而得到的, 而 (3.22, 3.23) 式则显然是从完全隐式格式出发得到的. 如果把 U^{n+1*} 看作是 u^{n+1} 的近似, 则不难证明 (3.22b) 式是与微分方程相容的, 且整个格式的截断误差为 $O(\Delta t+(\Delta x)^2)$. 这种做法也很容易推广到三维

$$(1-\mu_x\delta_x^2)U^{n+1*} = (1+\mu_y\delta_y^2+\mu_z\delta_z^2)U^n, \qquad (3.24a)$$

$$(1-\mu_y\delta_y^2)U^{n+1**} = U^{n+1*}-\mu_y\delta_y^2U^n, \qquad (3.24b)$$

$$(1-\mu_z\delta_z^2)U^{n+1} = U^{n+1**}-\mu_z\delta_z^2U^n. \qquad (3.24c)$$

而且前面刚提到的二维格式也适用于此.

对 (3.22) 、(3.23) 或 (3.24) 式做傅里叶分析, 不难证明这些格式与一维情形一样都是无条件稳定的. 不过, 不论从 (3.22) 式还是从 (3.23) 式的形式出发, 要应用最大值原理证明格式的稳定性就要困难得多.

另一种由俄罗斯学者提出并倍加推崇的方法是局部一维 (LOD) 格式, 这类格式每次

[1] D'yakonov, E.G. (1964), Difference schemes of second order accuracy with a splitting operator for parabolic equations without mixed partial derivatives, *Zh. Vychisl. Mat. i Mat. Fiz.*, **4**, 935-41.

[2] Douglas, J. Jr and Rachford, H.H. Jr (1956), On the numerical solution of the heat conduction problems in two and three variables, *Trans. Amer. Math. Soc.* **82**, 421-39.

只处理一个变量．把这种思想与 Crank-Nicolson 格式结合，并引入两个中间时间步，就得到

$$(1-\frac{1}{2}\mu_x\delta_x^2)U^{n+*} = (1+\frac{1}{2}\mu_x\delta_x^2)U^n, \tag{3.25a}$$

$$(1-\frac{1}{2}\mu_y\delta_y^2)U^{n+**} = (1+\frac{1}{2}\mu_y\delta_y^2)U^{n+*}, \tag{3.25b}$$

$$(1-\frac{1}{2}\mu_z\delta_z^2)U^{n+1} = (1+\frac{1}{2}\mu_z\delta_z^2)U^{n+**}. \tag{3.25c}$$

由于这些中间值与原微分方程毫不相容，处理其边界条件时需要特别加以注意．例如，记 V^n 、 V^{n+*} 为由 (3.25) 式的二维等价格式得到的数值解，和 Peaceman-Rachford 格式 (3.15) 式做比较．消去 V^{n+*} 后得到

$$(1-\frac{1}{2}\mu_y\delta_y^2)V^{n+1} = (1+\frac{1}{2}\mu_y\delta_y^2)(1-\frac{1}{2}\mu_x\delta_x^2)^{-1}(1+\frac{1}{2}\mu_x\delta_x^2)V^n, \tag{3.26}$$

将其与 (3.13) 式对比即知：若令

$$V^n = (1-\frac{1}{2}\mu_x\delta_x^2)U^n, \quad V^{n+*} = (1+\frac{1}{2}\mu_x\delta_x^2)U^n, \tag{3.27}$$

则 LOD 格式就等价于 Peaceman-Rachford 格式．充分利用这一关系就可以得到相当不错的边界条件．

3.4 曲线边界

在二维或更高维空间，另一个重要的新问题是区域通常更加复杂．在一维时只需考虑有限区间，该有限区间可以标准化为 [0, 1] 区间．到目前为止，我们所考虑的二维区域仅限于 [0, 1] 区间的自然推广，即单位正方形；而一般的二维区域会有曲线边界 (curved boundary), 且完全可能不是单连通的．一个简单的有物理背景的例子就是上面有一个圆洞的正方形平板．由此产生的一些困难主要是计算上的，关系到如何安排运算．例如，在一般的区域上给出一个正方形或矩形网格，则各网格线上的网格点数就不尽相同．本书不考虑这类问题，但我们必须讨论如何将边界条件吸收到有限差分方程中去．

我们以图 3-6 所示的问题为例． Ω 是以单位正方形四条边和圆心在 $(\frac{1}{2}, \frac{1}{2})$，半径为 0.33 的圆周 $\partial\Omega$ 为边界的区域，在 Ω 上考虑热传导方程

$$u_t = u_{xx} + u_{yy}. \tag{3.28}$$

在直线 $x+y=1$ 的上方区域边界的两个直线段和半圆弧上，边界条件给定为 $u=0$. 在

其余的边界上，即满足 $x+y<1$ 的边界上，边界条件给定为法向导数为零. 即

$$u(x,y,t)=0, \quad 在 \quad \{x+y\geqslant 1\}\cap\partial\Omega 上, \tag{3.29}$$

$$\frac{\partial u}{\partial n}=0, \quad 在 \quad \{x+y<1\}\cap\partial\Omega 上. \tag{3.30}$$

除位于东北 (NE) 和西南 (SW) 两角附近的两个对称小区域之外,初始条件给定为 $u(x,y,0)=0$. 物理上这个问题模拟的是有一个圆洞的正方形平板上的热流，平板的一半边界上温度维持在零度，另一半边界是绝热的 (insulated); 初始时刻，除了两个热点之外，平板上其他点的温度为零度.

如图 3-6 所示，正方形区域上给出了均匀分布的正方形网格，网格尺度为 $\Delta x=\Delta y=1/J$. 在该网格的许多网格点上可以使用标准的差分逼近，但对一些靠近边界的点就需要使用特殊的公式. 例如点 P, 其东西两侧的网格点都是正常的网格点，因此计算 $\delta_x^2 U$ 时不会有什么困难，但其南侧的网格点却在圆圈内. 图 3-7 将这一情形作了局部放大.

我们需要用 P, N 点以及边界上 B 点的函数值计算 u_{yy} 的近似值. 这种逼近很容易构造，其结果是标准三邻近点差分公式在非均匀分布网格点上的推广

有许多方法可以推导出所需的公式，最直接的方法之一是将 P 和 N 的中点 P_+ 以及

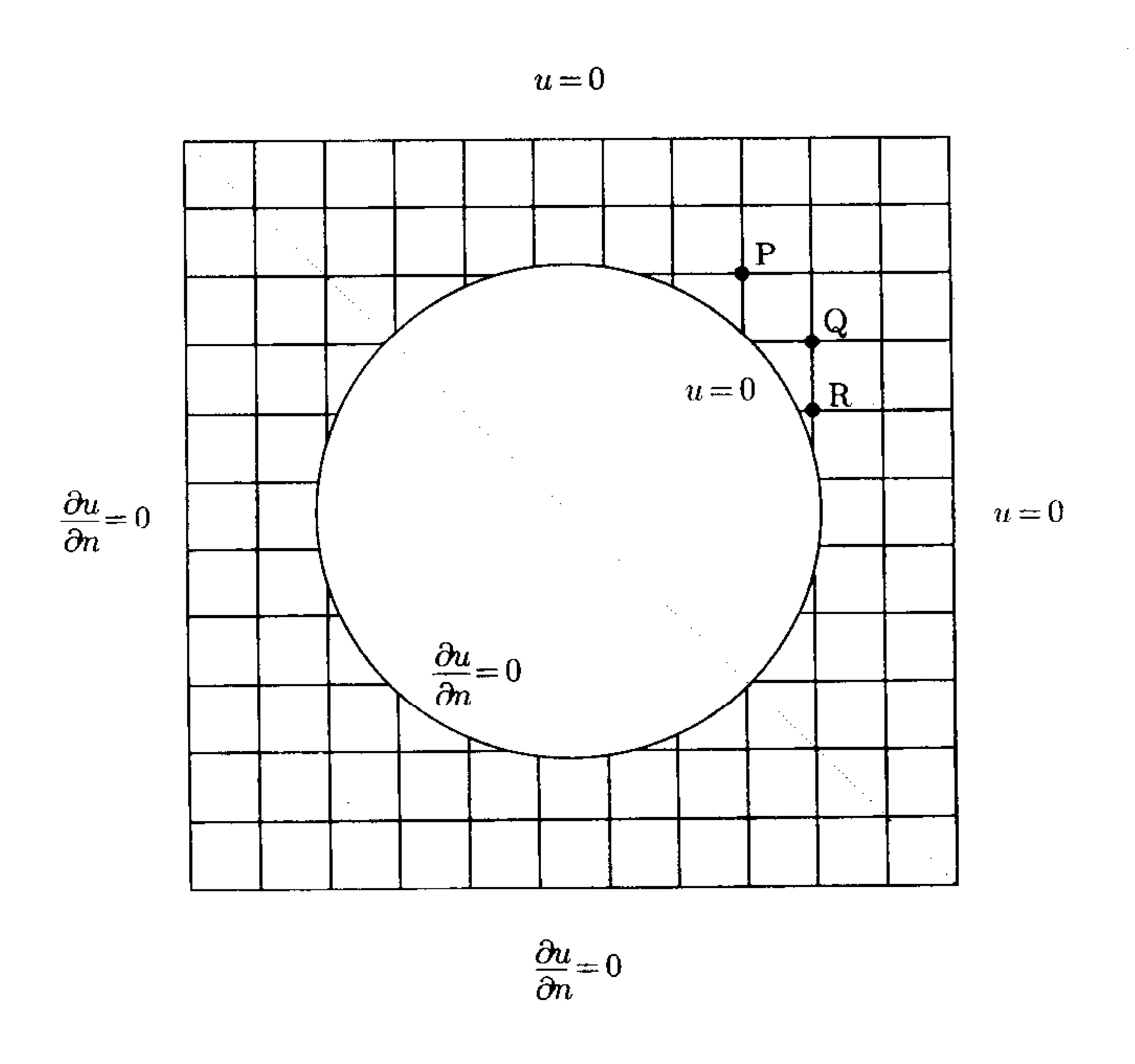


图 3-6 曲线边界的模型问题

P 和 B 的中点 P_- 这两点处的一阶偏导数近似地表示为

$$\frac{u_N-u_P}{y_N-y_P}\approx\frac{\partial u}{\partial y}(P_+),\quad \frac{u_P-u_B}{y_P-y_B}\approx\frac{\partial u}{\partial y}(P_-).$$

由于这两个中点间的距离是 $(y_N-y_B)/2$, 所以有所需的近似计算公式

$$u_{yy}\approx\frac{2}{y_N-y_B}\left(\frac{u_N-u_P}{y_N-y_P}-\frac{u_P-u_B}{y_P-y_B}\right).\tag{3.31a}$$

在这个例子中 $y_P-y_S=\Delta y$, 因此存在 $0<\alpha<1$ 使得 $y_P-y_B=\alpha\Delta y$. 所以近似计算公式变为

$$u_{yy}\approx\frac{2}{(\alpha+1)(\Delta y)^2}u_N-\frac{2}{\alpha(\Delta y)^2}u_P+\frac{2}{\alpha(\alpha+1)(\Delta y)^2}u_B.\tag{3.31b}$$

类似地，要计算 Q 这类点上导数 u_{xx} 的近似值，可以用 D 点代替相应的网格点，得到所需的近似计算公式. 还有一些像 R 点这样的点，这些点有两个邻近网格点在区域之外，因此 x 和 y 两个方向的导数都需要做特殊处理.

值得指出的是以上这些公式与利用内部及边界点向虚拟的外部点做外推 (extrapolating) 后，再应用标准差分格式得到的公式是一样的. 例如，利用二阶外推公式 (quadratic extrapolation)

$$u_S=\frac{\alpha(1-\alpha)u_N+2u_B-2(1-\alpha^2)u_P}{\alpha(\alpha+1)}\tag{3.32}$$

就可以得到 (3.31b) 式.

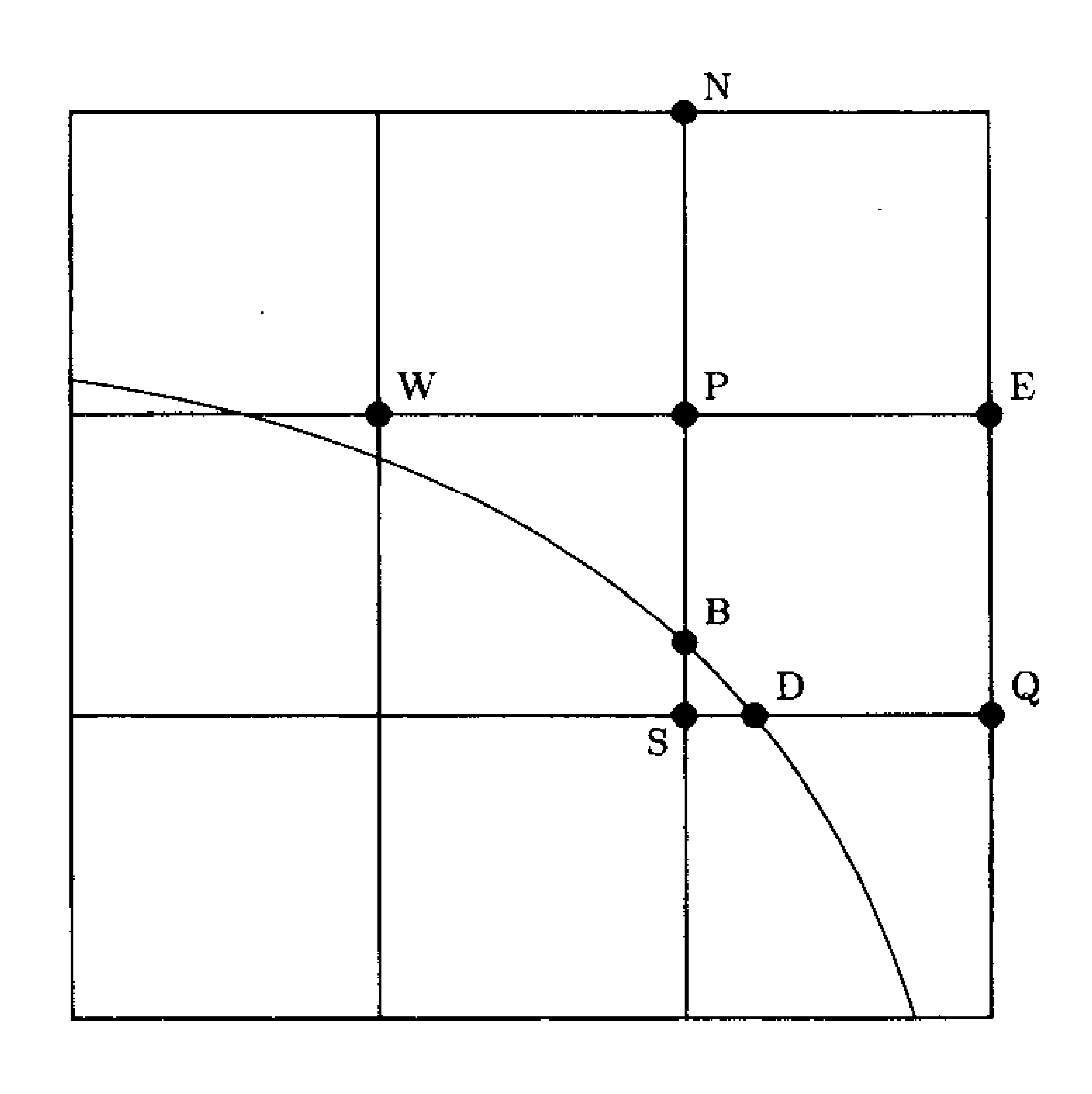


图 3-7 曲线边界上的狄利克雷边界条件

如果我们用的是显式方法，一般就会得到以下形式的差分格式

$$\begin{aligned}\frac{U_{r,s}^{n+1}-U_{r,s}^{n}}{\Delta t} &= \frac{2}{(1+\alpha)(\Delta x)^2}U_{r+1,s}^{n}+\frac{2}{\alpha(1+\alpha)(\Delta x)^2}u_B \\ &\quad +\frac{2}{(1+\beta)(\Delta y)^2}U_{r,s+1}^{n}+\frac{2}{\beta(1+\beta)(\Delta y)^2}u_D \\ &\quad -\left(\frac{2}{\alpha(\Delta x)^2}+\frac{2}{\beta(\Delta y)^2}\right)U_{r,s}^{n},\end{aligned} \tag{3.33}$$

其中 u_B 和 u_D 是边界上的给定值. 当时间步长 Δt 不太大时，以上用前一时间步的值给出的 $U_{r,s}^{n+1}$ 的表达式保持了一个重要的性质，即其表达式中各项的系数均非负. 显式格式的稳定性约束与标准情形类似， 误差分析也没什么两样. 但对时间步的约束很有可能比是矩形边界的时候更加苛刻了，因为此时的条件是

$$2\left(\frac{\mu_x}{\alpha}+\frac{\mu_y}{\beta}\right)\leqslant 1, \tag{3.34}$$

而 α 和 β 都是小于 1 的数，而且可能会很小. 例如，对正在讨论的这个问题，若取 $\Delta x=\Delta y=1/50$, 则允许的最大时间步为 0.000 008, 而对无洞问题所允许的最大时间步为 0.000 01, 可见差别不大. 但是，如果在 x 和 y 两个方向上同时将网格加密 1 倍，则无洞问题所允许的最大时间步减小到原来的 $\frac{1}{4}$, 而有洞问题所允许的最大时间步长则减小到原来的 $\frac{1}{800}$, 变成了 [1] 10^{-8}. 这类问题的存在要求我们使用隐式方法，或更合适的网格尺度，见下面的讨论.

边界上的法向导数可以做类似的处理， 当然要复杂得多. 例如，在图 3-8 中，假设给定了 B 点的外法向导数值， B 点的法线交水平网格线 WPE 于 Z 点，设相应线段的长度为

$$ZP=p\Delta x,\quad PB=\alpha\Delta y,\quad BZ=q\Delta y,$$

其中 $0\leqslant p\leqslant 1$, $0<\alpha\leqslant 1$, $0<q\leqslant\sqrt{1+(\Delta x)^2/(\Delta y)^2}$. 法向导数可以近似地表示为

$$\frac{u_B-u_Z}{q\Delta y}\approx\frac{\partial u}{\partial n}=g(B), \tag{3.35}$$

其中 u_Z 的值可以由 u_W 和 u_P 的线性插值得到

$$u_Z\approx p\,u_W+(1-p)u_P. \tag{3.36}$$

如前所述，我们可以用 u_W, u_P, u_E 来近似计算 u_{xx}, 用 u_S, u_P, u_B 来近似计算 u_{yy}, 消去 u_B 和 u_Z 就得到

---

[1] 原书中为 10^{-7}, 但显然 $0.000\,008\times 1/800=10^{-8}$.

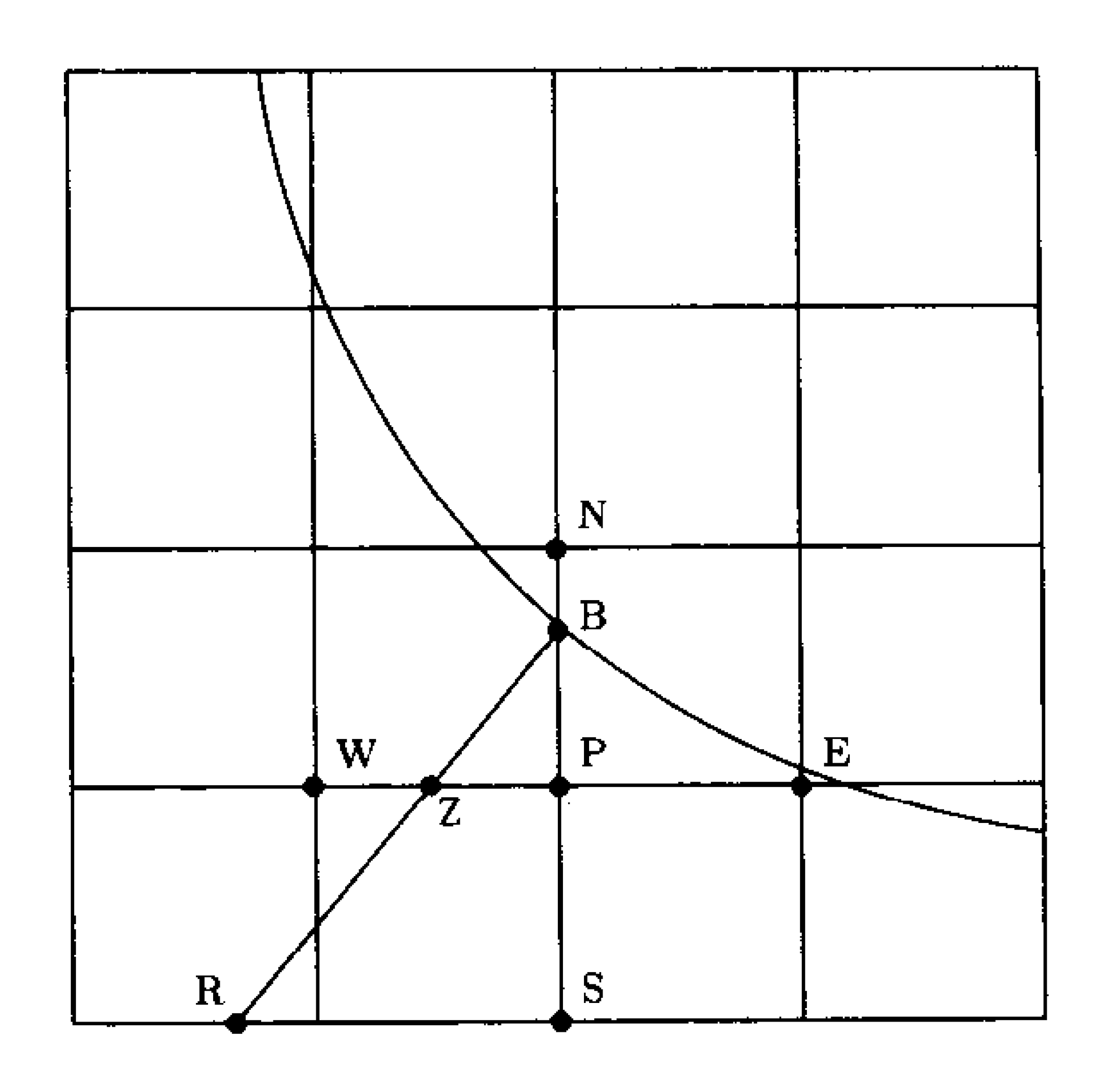


图 3-8 曲线边界上的诺伊曼 (Neumann) 条件

$$\begin{aligned}\frac{U_P^{n+1}-U_P^n}{\Delta t} &= \frac{U_E^n-2U_P^n+U_W^n}{(\Delta x)^2} \\ &\quad +\frac{1}{(\Delta y)^2}\left\{\frac{2}{\alpha(\alpha+1)}U_B^n-\frac{2}{\alpha}U_P^n+\frac{2}{\alpha+1}U_S^n\right\} \\ &= \frac{U_E^n-2U_P^n+U_W^n}{(\Delta x)^2}+\frac{1}{(\Delta y)^2}\left\{-\frac{2}{\alpha}U_P^n+\frac{2}{\alpha+1}U_S^n\right\} \\ &\quad +\frac{2}{\alpha(\alpha+1)(\Delta y)^2}\left\{pU_W^n+(1-p)U_P^n+qg_B\Delta y\right\} \\ &= \frac{1}{(\Delta x)^2}U_E^n+\left\{\frac{1}{(\Delta x)^2}+\frac{2p}{\alpha(\alpha+1)(\Delta y)^2}\right\}U_W^n \\ &\quad +\frac{2}{(\alpha+1)(\Delta y)^2}U_S^n-\left\{\frac{2}{(\Delta x)^2}+\frac{2(\alpha+p)}{\alpha(\alpha+1)(\Delta y)^2}\right\}U_P^n \\ &\quad +\frac{2q}{\alpha(\alpha+1)\Delta y}g_B. \end{aligned} \tag{3.37}$$

注意在该表达式中邻近点上 U 值的系数再次均为正数，同样的表达式也可以体现在 Crank-Nicolson 格式和 ADI 格式中．有了这种表达式，只要 Δt 适当的小，最大值原理就成立．将 (3.37) 式右端写成利用 U_N 值计算的标准格式，并由此推出 U_N 的外推公式，我们就会发现该格式与先用外推算出 U_N 再用标准格式计算是等价的．

各种形式的 ADI 方法在这方面几乎相同，在应用中主要的区别在于计算三对角矩阵靠近边界的元素时所需要的额外的计算量，另外，这些矩阵的大小是不一样的．

这些逼近曲线边界的方法给出的截断误差阶低于正常内点处的截断误差阶，特别是在需要计算法向导数的边界点上．恰如 2.13 节一维例子所示，当网格加密时这些截断误

差不一定趋向于零．而要构造具有所需性质的高阶逼近并非易事．我们前面仅用 U_B 和 U_Z 两点构造了最简单的法向导数近似值．现假设将 B 点的法线延伸到图 3-8 中的 R 点，其中 $ZR = (q/\alpha)\Delta y$. 则该法向导数的一个高阶近似值为

$$g_B = \frac{\partial u}{\partial n} \approx \frac{(1+2\alpha)u_B + \alpha^2 u_R - (1+\alpha)^2 u_Z}{(1+\alpha)q\Delta y}. \tag{3.38}$$

这时要算 u_R 的插值就不那么简单了，而且 u_Z 和 u_R 的系数的符号相反，所得到的格式不满足最大值原理．看来除非加以进一步的约束，否则这类高阶格式在应用中很有可能不会带来满意的效果．

图 3-9 到图 3-12 显示了应用 Peaceman-Rachford ADI 方法得到的典型计算结果．用的是 x 和 y 方向各 50 个网格点的一致网格．各图的尺度不尽相同，视点靠近东北 (NE) 方向，因此较小的峰值在前面； u 的两个初始峰值向整个区域扩散，变得越来越小，每个图的尺度都调到使最大值在图中的显示高度都相同． u 的实际最大值也在图的说明中给出．计算结果清晰地显示出其中一个峰值向边界外扩散，而另一个峰值则扩散到圆孔周围，因为热流不能穿过绝热的边界．最终热流绕过圆孔从另外一半边界传出．在图 3-11 中已经能开始看到这一点，而另一个峰值已经几乎完全消失．

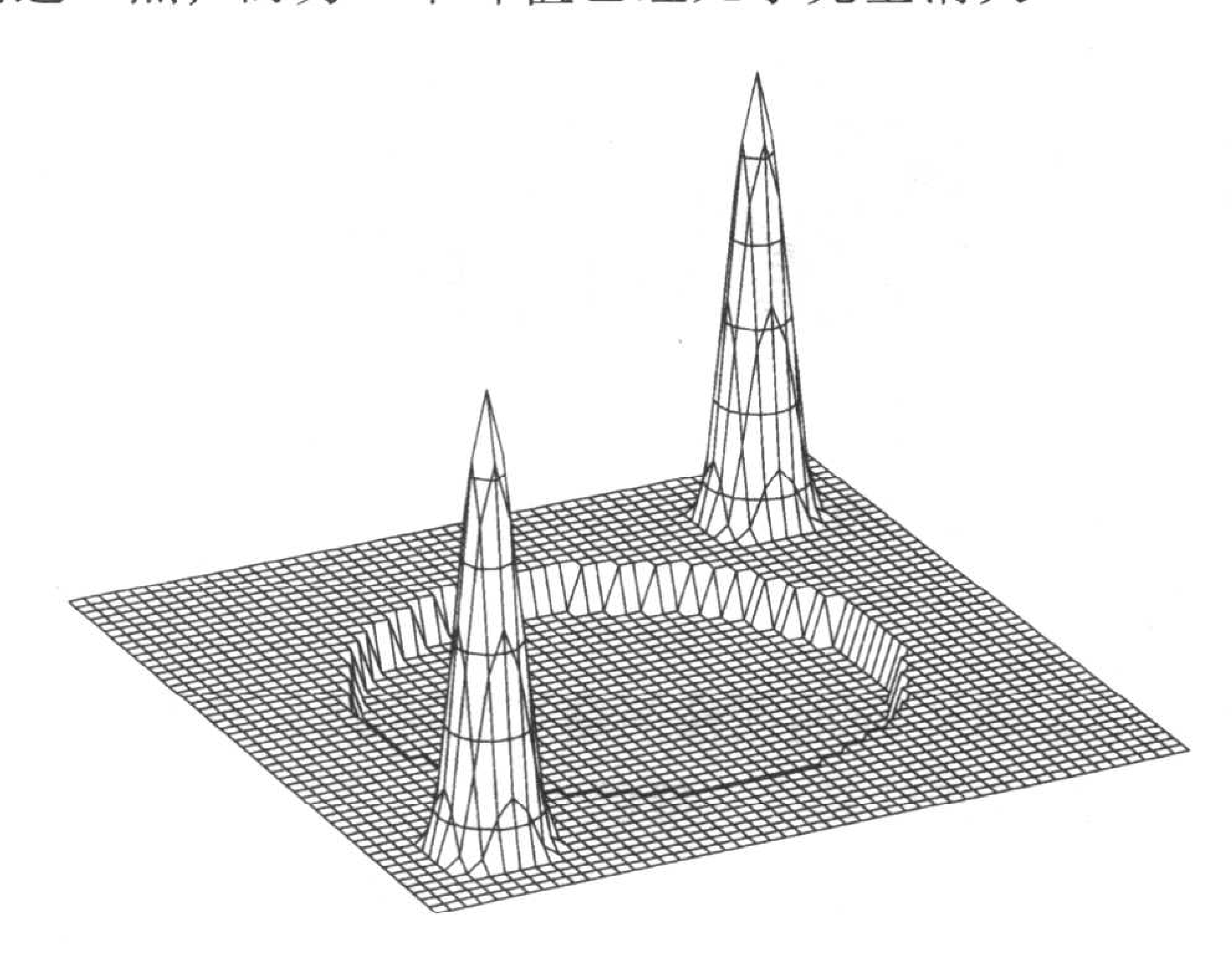


图 3-9　图 3-6 所示的区域上热流计算的初始值；最大值 $u = 1$

3.5　应用于一般抛物型问题

在高维时， (3.1) 式也会取以下形式

$$e\frac{\partial u}{\partial t} = \nabla\cdot(b\nabla u - \mathbf{a}u) + cu + d, \qquad \mathbf{x}\in\Omega, \tag{3.39}$$

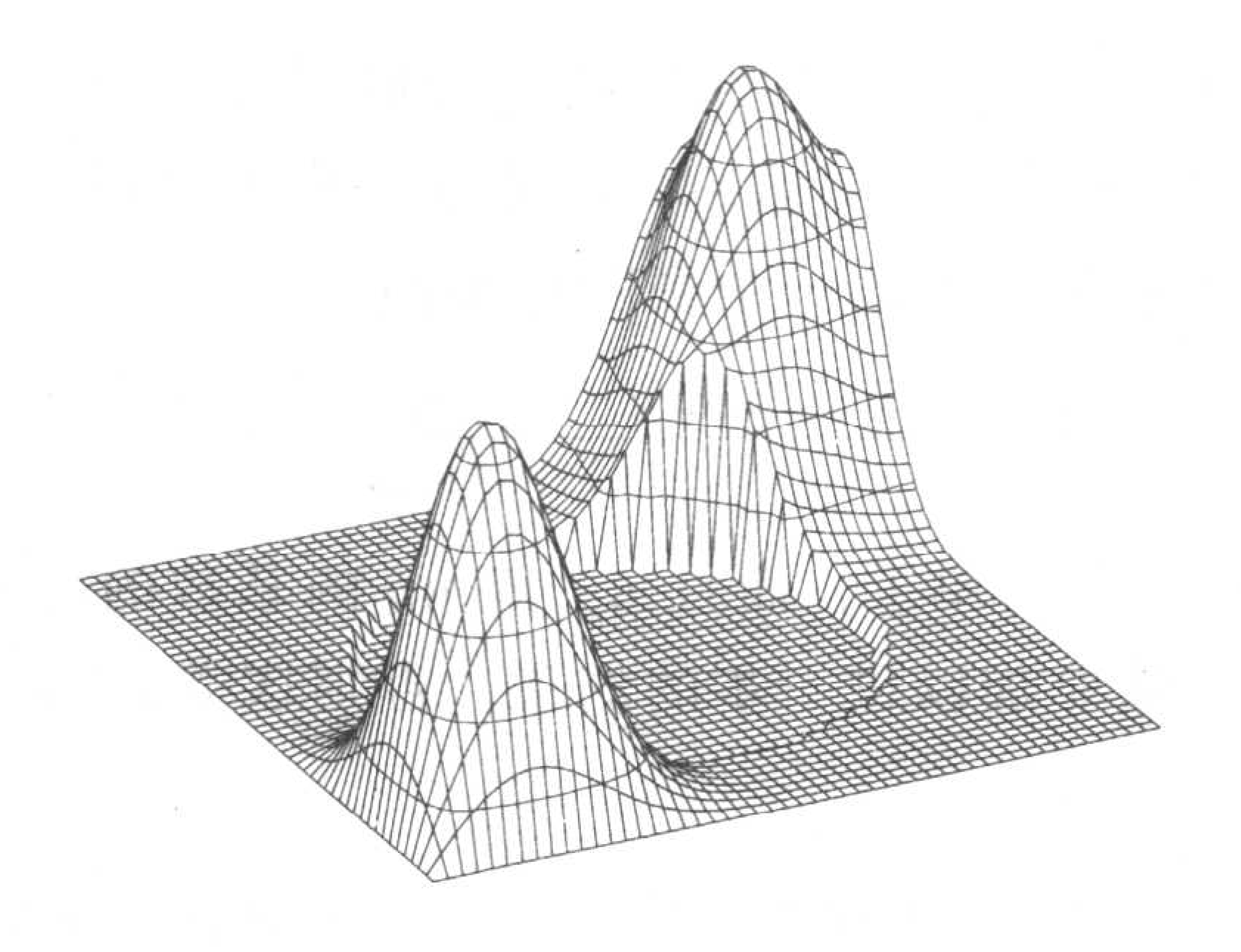


图 3-10　以图 3-9 为初始值 $t = 0.005$ 时的数值解；最大值 $U = 0.096$

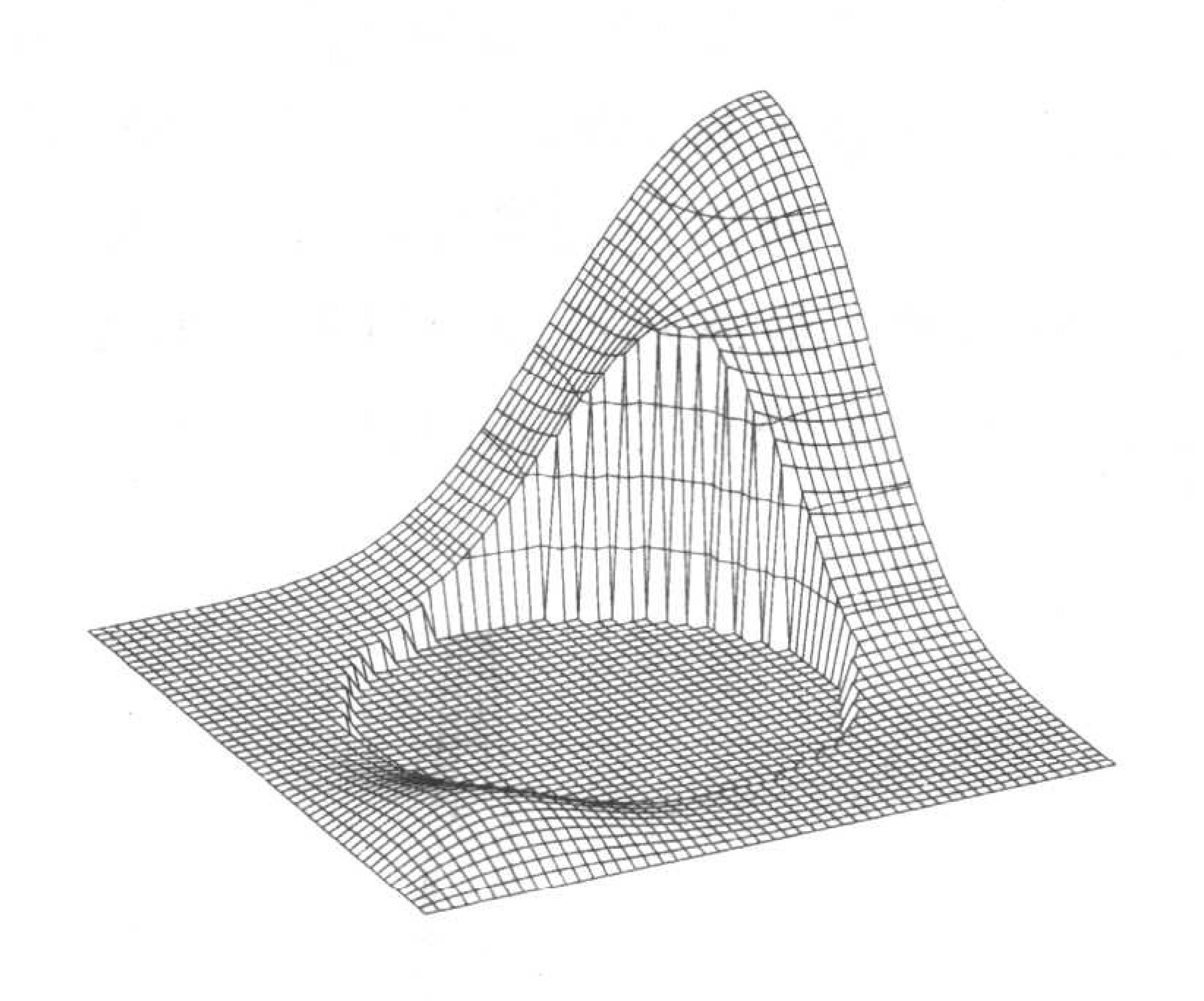


图 3-11　以图 3-9 为初始值 $t = 0.02$ 时的数值解；最大值 $U = 0.060$

其中所有的系数都依赖于 $\mathbf{x}$ 和 t, 对非线性问题还依赖于 u. Ω 可能不再是盒形的而是具有十分复杂形状的区域，边界条件具有一般的形式

$$\alpha_0 u + \alpha_1 \partial u/\partial n = \alpha_2, \qquad \mathbf{x} \in \partial\Omega, \tag{3.40}$$

其中 $\partial/\partial n$ 表示沿外法向求导，$\alpha_0 \geqslant 0, \alpha_1 \geqslant 0, \alpha_0 + \alpha_1 > 0$. 如第 1 章所述，我们在前面着重讨论的方法和思想对更一般的问题仍然适用. 这里我们仅就可能会出现的其他一些困难作一些评注.

在建立差分格式时通常仍选用规则网格 (regular mesh), 所用的格式与前面所讲的在本质上也是一样的，只是系数在局部取值. 主要的问题出在曲线边界上，这时边界条件必须用到非规则分布的点集；由此引起的网格尺度的差异和系数的变化就更突显出使用隐式格式的必要性.

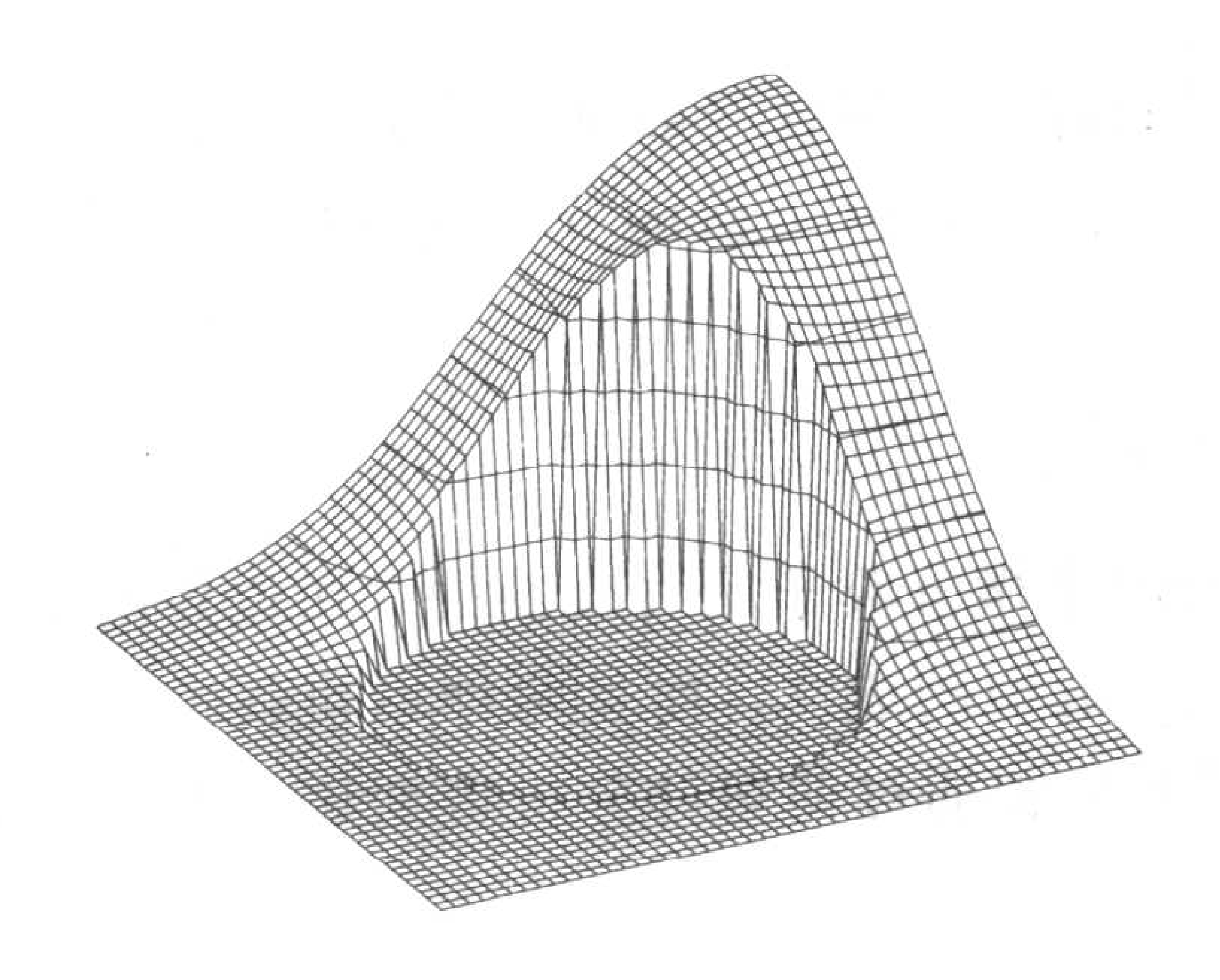


图 3-12 以图 3-9 为初始值 $t=0.05$ 时的数值解；最大值 $U=0.038$

另一种作法是选用能更好地反映边界的非规则网格 (irregular mesh). 求解 (2.156) 式的圆柱对称格式的技巧可以推广到弯曲边界，一族弯曲的坐标曲线大致平行于边界，另一族则大致垂直于边界，这就导出了所谓“边界 - 匹配网格 (boundary-fitted mesh)”和有限体积法. 还有一种作法是采用完全非结构的三角形网格和有限元方法. 这两种做法都超出了本书的范围，不过第 4 章和第 5 章将会讨论有限体积法，第 6 章将会讨论有限元方法.

我们所讨论的格式都是两时间层格式，在大规模问题的计算中这类格式的优越性体现在其所需计算机存储空间最小，大多数格式所占的存储空间比单个时间层上 U 所占的存储空间大不了多少. 求解矩形网格上的差分方程可以应用 ADI 或 LOD 方法，只需反复调用三对角方程解法器. 当然，为了得到精确的边界条件，我们仍需在边界上做些调整，特别是在内凹的角点处. 以上都假设方程是线性的，当方程的非线性不太强时，可将前一时间层线性化，这样仍可以使用隐式格式计算，并且仍有望达到合理的精度. 到目前为止，除非在常系数线性格式、一致网格的矩形区域和周期边界条件等条件下，我们无法利用完全的傅里叶分析方法研究算法的分析性质. 不过，可以证明由局部傅里叶分析得到的稳定性条件仍然是必要条件，本质上这是因为在各种情形下最不稳定的波型都是高频的，因而是局部增长的. 对线性问题，先将差分方程的系数在局部取定为常数，然后代入傅里叶波型算出不稳定的解；对非线性问题，可以对线性化的问题做同样的分析. 第 5 章中我们将进一步讨论有关问题.

对于线性问题，证明收敛性的最得力的工具是最大值原理，只要正确地选择了差分格式，这一点对边界条件含有导数的问题尤为重要，则最大值原理依然成立. 不过，必须注意，在证明收敛性时所需的条件和证实会出现不稳定现象时所需的条件之间一般会

存在缺口. 这一点当边界条件为一般含导数项的边界条件，且边界为弯曲的情况时尤为突出. 特别地，对于这类问题， ADI 方法的分析变得更加复杂. 到目前为止，我们所依赖的是以下因式分解等式：

$$(1-\frac{1}{2}\mu_x\delta_x^2)(1-\frac{1}{2}\mu_y\delta_y^2)U=(1-\frac{1}{2}\mu_y\delta_y^2)(1-\frac{1}{2}\mu_x\delta_x^2)U\ .$$

该等式之所以成立是因为两个微分算子是可交换的. 在边界附近的非规则分布的一组网格点上，这种因式分解只是近似的成立，而且两种不同顺序的因式分解所产生的余项是不同的. 因此，尽管前面我们曾通过傅里叶分析得到这类方法的重要性质，但在这里傅里叶分析的方法不再适用.

在过去的二三十年中，正如在第 7 章中将要讨论的，求解大型代数方程组的方法和软件取得了巨大的进展，尤其是继 Brandt[1] 工作之后 *多重网格法* (multigrid method) 的发展. 这些进展为用隐式格式求解抛物型问题提供了一个新的视角，即只需依次求解一系列的椭圆问题，并由前一时间步提供所需数据. 因此本章的问题也可以用第 6 章和第 7 章所讲的方法求解. 这时 ADI 方法和 LOD 方法在作为代数解法器的预条件方面基本上不起作用，或只起十分次要的作用. 如果用有限元方法，则意味着变分原理将为算法提供分析框架.

到目前为止，我们还没有讨论过形如 (3.39) 式方程的特点. 方程中含 **a** 的项是对流项，与主项的扩散作用相比，该项的引入常常降低了系统的稳定性，这点我们将在后面两章及 6.8 节中进行研究. 我们假设了扩散项的系数 b 为标量，这也部分地说明了为什么那么多的方法和分析能够容易地进行推广. 我们可以引入对角张量，从而使沿各坐标方向的扩散系数互不相同，而且这也不会带来多大变化；不过在实际应用中这种各向异性不太可能恰巧与所选的坐标轴很好地吻合，因为各向异性反映的是真实的，比如在成层岩中的物理效果. 因此扩散张量有非对角项，这时方程就会有混合二阶导数项，从而一般可表示为

$$eu_t=b_{11}u_{xx}+2b_{12}u_{xy}+b_{22}u_{yy}-a_1u_x-a_2u_y+cu+d, \tag{3.41}$$

这样一来二维差分格式就会用到 9 个而不是原来的 5 个点，而三维差分格式就会用到 27 个点.

这会带来极大的困难，深入讨论这一问题超出了本书的范围. 我们在此仅就九点差分格式作些评注. 假设在正方形网格上用差商 $\Delta_{0x}\Delta_{0y}U/(\Delta x)^2$ 逼近混合导数 u_{xy}, 其结果是在东北 (NE) 、西北 (NW) 、东南 (SE) 或西南 (SW)(图 3-7 和图 3-8 中的指南针记号) 诸点中总会有一点的系数是负的，因而造成无法应用最大值原理. 不过， u_{xx} 和 u_{yy} 可以用对角二阶差分逼近，这会使这些系数成为正值. 例如，若 $b_{11}>b_{22}$, 则可以用两个对角

[1] Brandt, A. (1977) Multilevel adaptive solutions to boundary-value problems. *Math. Comp.* **31**, 333-90.

方向的二阶差分近似计算 $b_{22}(u_{xx}+u_{yy})$，于是得到的正数为 $b_{22}/2(\Delta x)^2$. 如果 $b_{22} \geqslant |b_{12}|$，这就保证了最大值原理成立. 不过要注意，这并不总是可行的，因为只要 $b_{11}b_{22} > b_{12}^2$，则 (3.41) 式中的算子就是椭圆型的. 这说明实际中最困难的情况是某方向的扩散系数比其他方向的大得多.

文献注记与推荐读物

第 2 章所推荐的进一步阅读的文献也完全适用于本章，同样适用于本章的还有第 6 章中为椭圆型问题，第 7 章中为求解由椭圆型问题离散化得到的线性代数方程组所推荐的大部分读物.

Mitchell 和 Griffiths (1980) 的书中详尽地讨论了 ADI 及相关的方法； Yanenko (1971) 的书中相当全面地介绍了俄罗斯作者们对该领域的许多贡献.

习　题

3.1　证明对于方程 $u_t = b(u_{xx}+u_{yy})$ 的解，其中 $b=b(x,y)>0$，显式格式

$$\frac{U^{n+1}-U^n}{\Delta t} = b\left[\frac{\delta_x^2 U^n}{(\Delta x)^2} + \frac{\delta_y^2 U^n}{(\Delta y)^2}\right]$$

的截断误差主项为

$$T^n = \frac{1}{2}\Delta t u_{tt} - \frac{1}{12}b\left[(\Delta x)^2 u_{xxxx} + (\Delta y)^2 u_{yyyy}\right].$$

若在矩形区域 $[0,\ X]\times[0,\ Y]$ 的边界的所有点上，对所有 $t>0$，$u(x,y,t)$ 的值为已知，设问题的解充分光滑，证明当 $b\Delta t\left[(\Delta x)^{-2}+(\Delta y)^{-2}\right] \leqslant \frac{1}{2}$ 时，误差上界在 $0 \leqslant t \leqslant t_F$ 的范围内满足估计式

$$\left|U_{r,s}^n - u(x_r,y_s,t_n)\right| \leqslant t_F\left\{\frac{1}{2}\Delta t M_1 + \frac{1}{12}\left[(\Delta x)^2 M_2 + (\Delta y)^2 M_3\right]\right\},$$

并确认常数 M_1, M_2 和 M_3.

3.2　证明 Peaceman-Rachford ADI 方法关于方程 $u_t = \nabla^2 u$ 的截断误差主项为

$$\begin{aligned} T^{n+1/2} \quad &= \quad (\Delta t)^2\left[\frac{1}{24}u_{ttt} - \frac{1}{8}(u_{xxtt}+u_{yytt}) + \frac{1}{4}u_{xxyyt}\right] \\ &\quad -\frac{1}{12}\left[(\Delta x)^2 u_{xxxx} + (\Delta y)^2 u_{yyyy}\right]. \end{aligned}$$

如果引入扩散系数变量 b，上式会有什么变化？

3.3 证明三维热传导方程 $u_t = u_{xx} + u_{yy} + u_{zz}$ 的 Douglas-Rachford 格式

$$\begin{aligned}(1-\mu_x\delta_x^2)U^{n+1*} &= (1+\mu_y\delta_y^2+\mu_z\delta_z^2)U^n,\\ (1-\mu_y\delta_y^2)U^{n+1**} &= U^{n+1*}-\mu_y\delta_y^2U^n,\\ (1-\mu_z\delta_z^2)U^{n+1} &= U^{n+1**}-\mu_z\delta_z^2U^n\end{aligned}$$

在盒形区域上是无条件稳定的.

3.4 设在有曲线边界的二维区域上，用显式方法在一致网格上求解热传导方程 $u_t = u_{xx} + u_{yy}$, 设在所有边界点上给定狄利克雷边界条件. 在邻近边界的节点上格式被修正为

$$\begin{aligned}\frac{U_{r,s}^{n+1}-U_{r,s}^n}{\Delta t} = &\frac{2}{(1+\alpha)(\Delta x)^2}U_{r+1,s}^n+\frac{2}{\alpha(1+\alpha)(\Delta x)^2}U_B^n+\frac{2}{(1+\beta)(\Delta y)^2}U_{r,s+1}^n\\ &+\frac{2}{\beta(1+\beta)(\Delta y)^2}U_D^n-\left(\frac{2}{\alpha(\Delta x)^2}+\frac{2}{\beta(\Delta y)^2}\right)U_{r,s}^n,\end{aligned}$$

其中 B 和 D 是邻近 $(x_r,\ y_s)$ 的边界点，它们到 $(x_r,\ y_s)$ 点的距离分别为 $\alpha\Delta x$ 和 $\beta\Delta y$. 试给出在该点上截断误差的主项，并证明在所有点上都有如下的误差界

$$|T_{r,s}^n| \leqslant T := \frac{1}{2}\Delta t M_{tt}+\frac{1}{3}(\Delta x M_{xxx}+\Delta y M_{yyy})+\frac{1}{12}[(\Delta x)^2M_{xxxx}+(\Delta y)^2M_{yyyy}].$$

试叙述最大值原理成立时要求 Δt 满足的约束条件，并证明当该条件满足时，数值解在 $0 \leqslant t \leqslant t_F$ 范围内的误差界为 $t_F T$.

3.5 设在正方形区域 $0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant 1$ 上，用显式方法在一致网格上求解热传导方程 $u_t = u_{xx} + u_{yy}$. 设在正方形的 $x = 0$ 的边上给定边界条件 $u_x = 0$, 设在其余边界点上给定狄利克雷边界条件. 设在 $x = 0$ 的网格点旁又增加了一列 $x = -\Delta x$ 的网格点，而相应的额外增加出来的未知量 $U_{-1,s}^n$ 可利用边界条件消去. 这样处理后，在 $x = 0$ 的边界上格式就变为

$$\frac{U_{0,s}^{n+1}-U_{0,s}^n}{\Delta t} = \frac{2}{(\Delta x)^2}(U_{1,s}^n-U_{0,s}^n)+\frac{1}{(\Delta y)^2}\delta_y^2U_{0,s}^n.$$

证明这种网格点上的截断误差主项为

$$T_{0,s}^{n*} = \frac{1}{2}\Delta t\, u_{tt}-\frac{1}{12}\left[(\Delta x)^2u_{xxxx}+(\Delta y)^2u_{yyyy}\right]-\frac{1}{3}\Delta x u_{xxx},$$

并进一步证明当通常的稳定性条件满足时，数值解的误差满足

$$e_{r,s}^n \equiv U_{r,s}^n - u(x_r, y_s, t_n) = O(\Delta t)+O(\Delta x)+O((\Delta y)^2).$$

3.6 设在正方形一致网格上，用基于 Crank-Nicolson 格式的分数步方法求解矩形区域上的扩散方程 $u_t = b\nabla^2 u$. 证明当 b 为常数时， Peaceman-Rachford ADI 方法完全等价于 LOD 方法

$$\begin{aligned}(1-\frac{1}{2}\mu\delta_y^2)V^{n+\frac{1}{2}} &= (1+\frac{1}{2}\mu\delta_y^2)V^n,\\ (1-\frac{1}{2}\mu\delta_x^2)V^{n+1} &= (1+\frac{1}{2}\mu\delta_x^2)V^{n+\frac{1}{2}},\end{aligned}$$

其中 $\mu = b\Delta t/(\Delta x)^2$. 当 b 为 (x, y) 的函数时，记 U^n 为 ADI 格式算出的数值解，令 $V^0 = (1 - \frac{1}{2}\mu\delta_y^2)U^0$，证明两种格式仍然是相互关联的，并找出在整时间步和中间步处 U 和 V 之间的关系.

第 4 章　一维双曲型方程

4.1　特征线方法

线性对流方程

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \tag{4.1}$$

肯定是最简单的偏微分方程了. 但在固定的 (x,t) 网格上很好地逼近其解并非易事，现在这仍是许多数值分析文献热烈讨论的对象. 正如所知，线性对流方程是双曲型方程，有一族特征线 (characteristics)， u 沿着每条特征线都是常数，由此可以求出其精确解. 而特征线 是常微分方程

$$\frac{\mathrm{d}x}{\mathrm{d}t} = a(x,t), \tag{4.2a}$$

的解，沿每条特征曲线， $u(x,t)$ 满足

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} = 0. \tag{4.2b}$$

因此，由初始条件

$$u(x,0) = u^0(x), \tag{4.3}$$

其中 $u^0(x)$ 为已知函数，我们可以这样来构造方程的近似解：如图 4-1 所示， 选取一组适

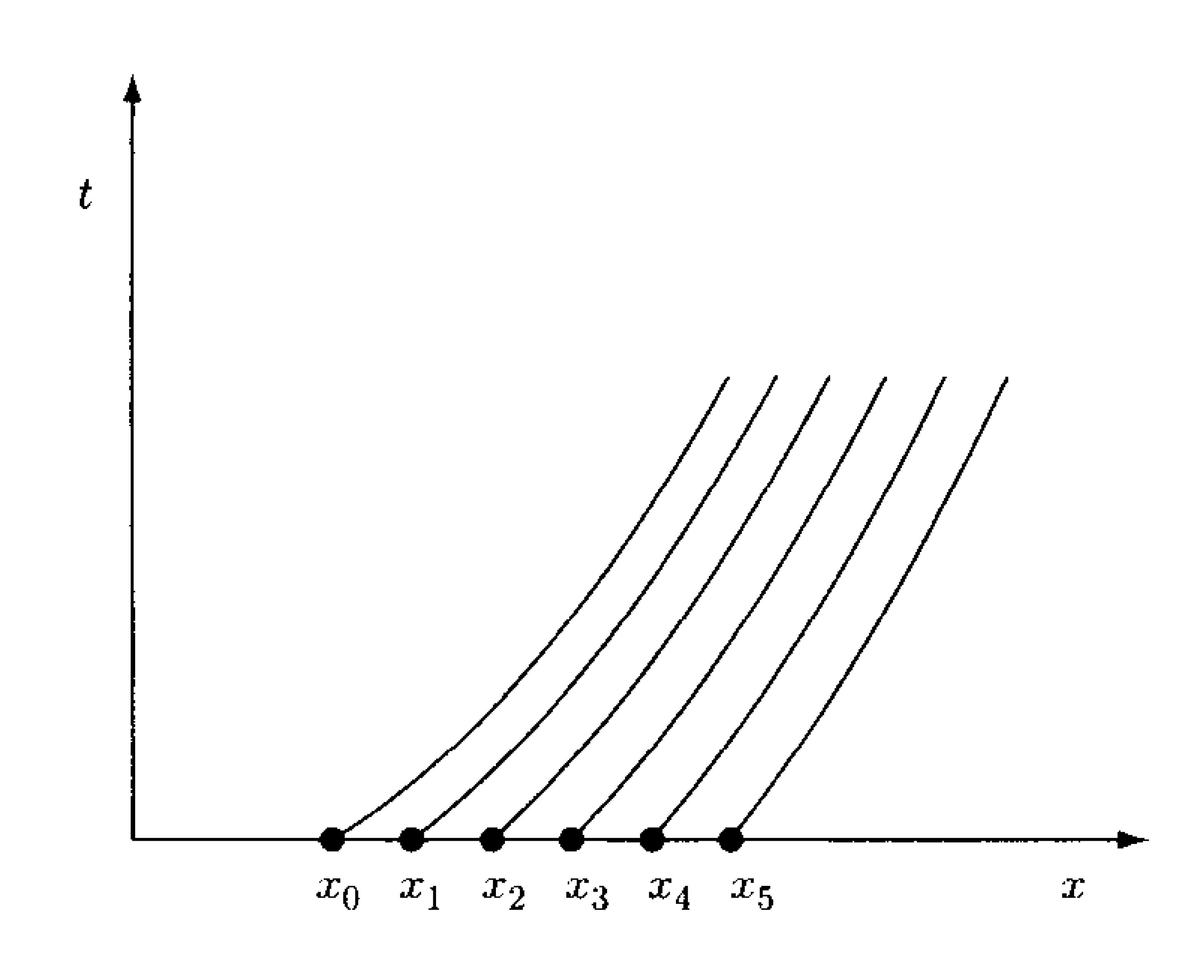


图 4-1　方程 $u_t + a(x,t)u_x = 0$ 的几条典型的特征线

当的点 $x_0, x_1, \ldots$，求解带有初始条件 $x(0) = x_j$ 的常微分方程 (4.2a) 的数值解，得到通过点 $(x_j, 0)$ 的特征线，在这些特征线上令 $u(x,t) = u^0(x_j)$．这即所谓的 特征线方法(method of characteristics)．注意对于该线性问题，$a(x,t)$ 是已知函数，因此只要 a 关于 x 利普希茨连续，关于 t 连续，特征线就不会相交．

若 a 是常数, 特征线方法将十分简单. 因为此时特征线是一族平行直线 $x - at =$ 常数，解可简单地表示为

$$u(x,t) = u^0(x - at). \tag{4.4a}$$

对非线性问题，若 a 仅是 u 的函数，即 $a = a(u)$, 则由于 u 沿着每条特征线为常数，特征线将仍是直线，但这些直线不再平行．因此我们仍然可以把解表示为如下形式

$$u(x,t) = u^0(x - a(u(x,t))t), \tag{4.4b}$$

不过到了某个时刻，特征线开始出现缠绕或相交时，上述形式的解就不再正确 (参见第 4.6 节)．

在任何发展和研究双曲型方程或方程组的数值方法时，特征线是至关重要的，下面我们会经常提到它. 我们将会考虑以下形式的 守恒律方程组(systems of conservation laws)

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{u})}{\partial x} = 0, \tag{4.5}$$

其中 $\mathbf{u} = \mathbf{u}(x,t)$ 是向量值未知函数，$\mathbf{f}(\mathbf{u})$ 是向量值 通量函数 (flux function)． 例如，若 $\mathbf{u}$ 有两个分量 u 和 v，$\mathbf{f}$ 有两个分量 $f(u,v)$ 和 $g(u,v)$，则 (4.5) 式的分量形式可表示为

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u,v) = 0, \tag{4.6a}$$

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x} g(u,v) = 0, \tag{4.6b}$$

或写成矩阵形式

$$\begin{pmatrix} \dfrac{\partial u}{\partial t} \\ \dfrac{\partial v}{\partial t} \end{pmatrix} + \begin{pmatrix} \dfrac{\partial f}{\partial u} & \dfrac{\partial f}{\partial v} \\ \dfrac{\partial g}{\partial u} & \dfrac{\partial g}{\partial v} \end{pmatrix} \begin{pmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial v}{\partial x} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.7}$$

若由 f 的偏导数定义雅可比矩阵

$$A(\mathbf{u}) := \frac{\partial \mathbf{f}}{\partial \mathbf{u}}, \tag{4.8}$$

则方程组亦可表示为

$$\mathbf{u}_t + A(\mathbf{u})\mathbf{u}_x = \mathbf{0}, \tag{4.9}$$

其特征速度 (characteristic speed) 是 A 的特征值. 若 A 的特征值都是实数且其所有特征向量张成整个欧氏空间，则称此方程组为双曲型的. 记 Λ 为所有特征值组成的对角矩阵，

$S=S(\mathbf{u})$ 为左特征向量组成的矩阵，于是

$$SA=\Lambda S. \tag{4.10}$$

在方程 (4.9) 两端左乘矩阵 S，得到方程组的 正规特征形式 (characteristic normal form)

$$S\mathbf{u}_t+\Lambda S\mathbf{u}_x=\mathbf{0}, \tag{4.11}$$

若存在 黎曼不变量 (Riemann invariant) $\mathbf{r}=\mathbf{r}(\mathbf{u})$，即满足 $\mathbf{r}_t=S\mathbf{u}_t$ 和 $\mathbf{r}_x=S\mathbf{u}_x$ 的向量，则上述方程组可表示为

$$\mathbf{r}_t+\Lambda\mathbf{r}_x=\mathbf{0}, \tag{4.12}$$

这是标量方程的直接推广，而标量方程的解可由 (4.4b) 给出．但这里 Λ 的分量一般依赖于 $\mathbf{r}$ 的所有分量，故特征线一般为曲线．另外，虽然对两个方程组成的方程组总可定义其黎曼不变量，但对更多方程的方程组，有时其黎曼不变量并不存在．

为了对 (4.5) 这类特征速度依赖于解的问题应用特征线方法，我们必须联立求解特征线方程及微分方程的正规特征形式 (4.11)，一步步向前计算．这显然是十分复杂的过程，但该方法却或许是数值求解此类方程组的最精确的方法．

二维双曲方程的特征曲面和解的形式极其繁杂多样，直接应用特征线方法将变得异常复杂．因此，尽管本章只考虑一维问题，为了与第 1 章所给出的总体思路相一致，我们不再对特征线方法做更详细的讨论．我们将仅考虑在固定的空间网格上设计的数值方法，时间步的长度在不同的时间步可以不同，但在所有空间点上同一时间步的步长必须一致．下面先讨论均匀网格上的显式方法．

4.2 CFL 条件

在 1928 年的一篇关于偏微分方程差分方法的重要论文 [1] 中，Courant，Friedrichs 和 Lewy 提出了通过依赖区域 (domain of dependence) 的概念判断差分方法收敛性的一个必要条件，这就是著名的 *CFL* 条件(CFL condition)．先考虑最简单的模型问题 (4.1)，这里 a 是正常数，正如所知，其解是 $u(x,t)=u^0(x-at)$，其中 u^0 是初始条件．沿经过点 (x_j,t_n) 的特征线找到其与初始直线的交点 $Q\equiv(x_j-at_n,0)$，由此可得到点 (x_j,t_n) 处的解 (见图 4-2)．

现假定我们采用显式格式

$$\frac{U_j^{n+1}-U_j^n}{\Delta t}+a\frac{U_j^n-U_{j-1}^n}{\Delta x}=0. \tag{4.13}$$

来计算有限差分逼近值，那么可以计算新的时间层上的值如下

[1] Courant, R.,Friedrichs, K.O. and Lewy, H. (1928), Über die partiellen Differenzengleichungen der mathematischen Physik, *Math. Ann.* **100**, 32-74.

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{a\Delta t}{\Delta x}(U_j^n - U_{j-1}^n) \\ &= (1-\nu)U_j^n + \nu U_{j-1}^n, \end{aligned} \tag{4.14}$$

其中

$$\nu = \frac{a\Delta t}{\Delta x}. \tag{4.15}$$

U_j^{n+1} 的值依赖于 U 在前一时间层两点上的值，而这两点的值又分别依赖于 t_{n-1} 时间层

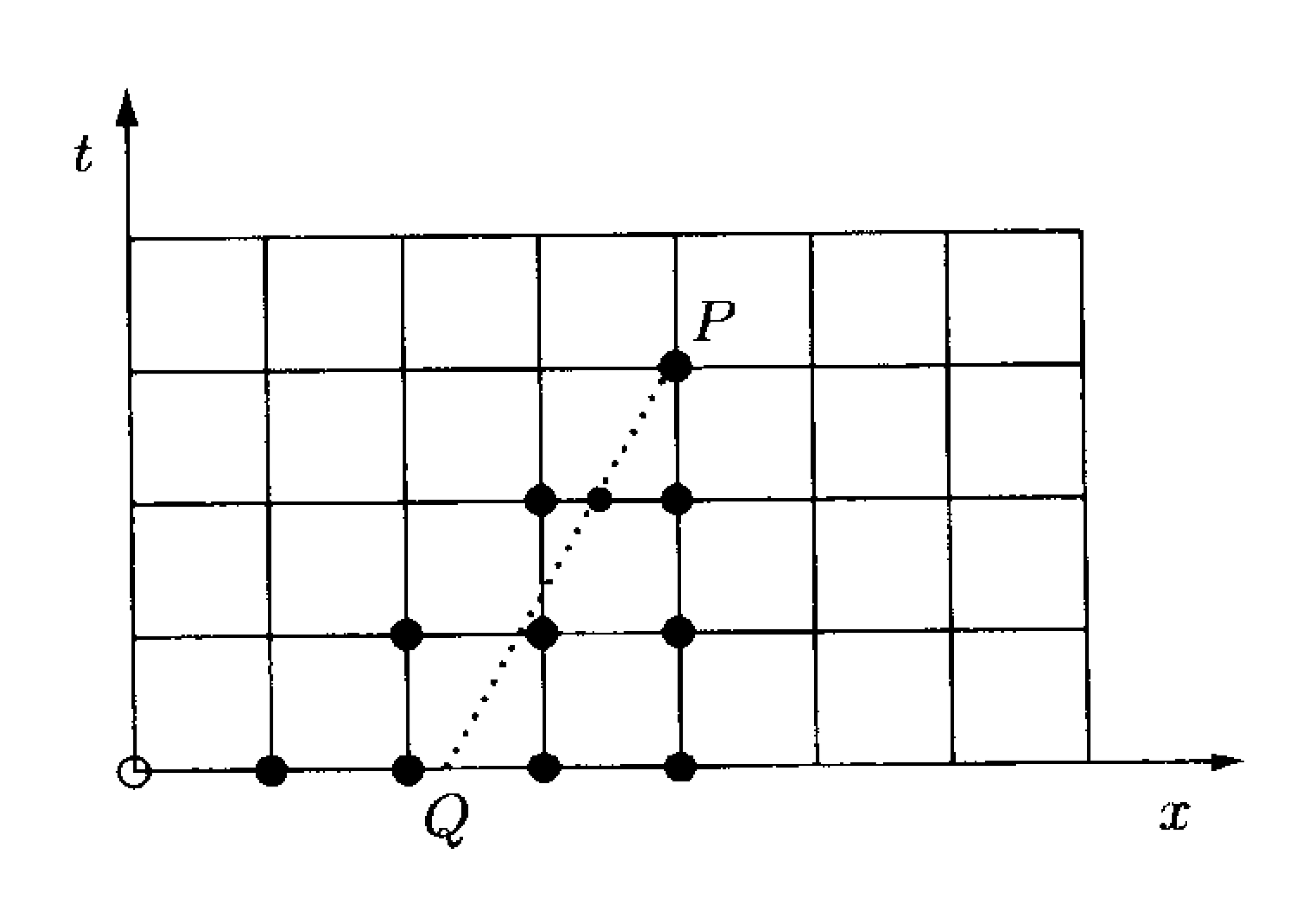


图 4-2　典型的依赖区域

上某两点的值，依此类推. 如图 4-2 所示，值 U_j^{n+1} 由以 (x_j, t_{n+1}) 为顶点的一个三角形中节点上的值决定，并最终由初始直线上的点

$$x_{j-n-1}, x_{j-n}, \cdots, x_{j-1}, x_j.$$

上的值所决定. 对于非齐次方程，要用源项 h_j^n 代替 (4.13) 右端的零，此时 U_j^{n+1} 依赖于三角形中所有点的值. 这个三角形就叫作该格式关于 U_j^{n+1} 或者点 (x_j, t_{n+1}) 的依赖区域(domain of dependence) .

微分方程在 (x_j, t_{n+1}) 点的依赖区域是指特征曲线从该点返回到初始直线的轨迹，这是因为对于非齐次方程 $u_t + au_x = h$ ，右端项 $h(x,t)$ 要在包括初始时刻点 $x = x_j - at_{n+1}$ 在内的整条特征曲线上取值. CFL 条件就是指，为保证格式收敛，偏微分方程的依赖区域必须包含在数值格式的依赖区域内.

图 4-3 显示了两种不满足 CFL 条件的情形. 两条特征线 PQ 和 PR 都处在格式的三角形依赖区域外. 在保持网格比 $\Delta t/\Delta x$ 不变的情况下加密网格，三角形依赖区域将保持不变. 若改变初始直线 $t=0$ 上 Q 点的一个小邻域内的初始条件，由于方程的解沿特征线是常数，微分方程在 P 处的解就要相应改变. 但由于用来计算数值解的数据没有改变，因此 P 点的数值解就不会改变. 这样， P 点的数值解就不可能收敛到所需的结果. 当特征线为 RP 时，可做类似的讨论.

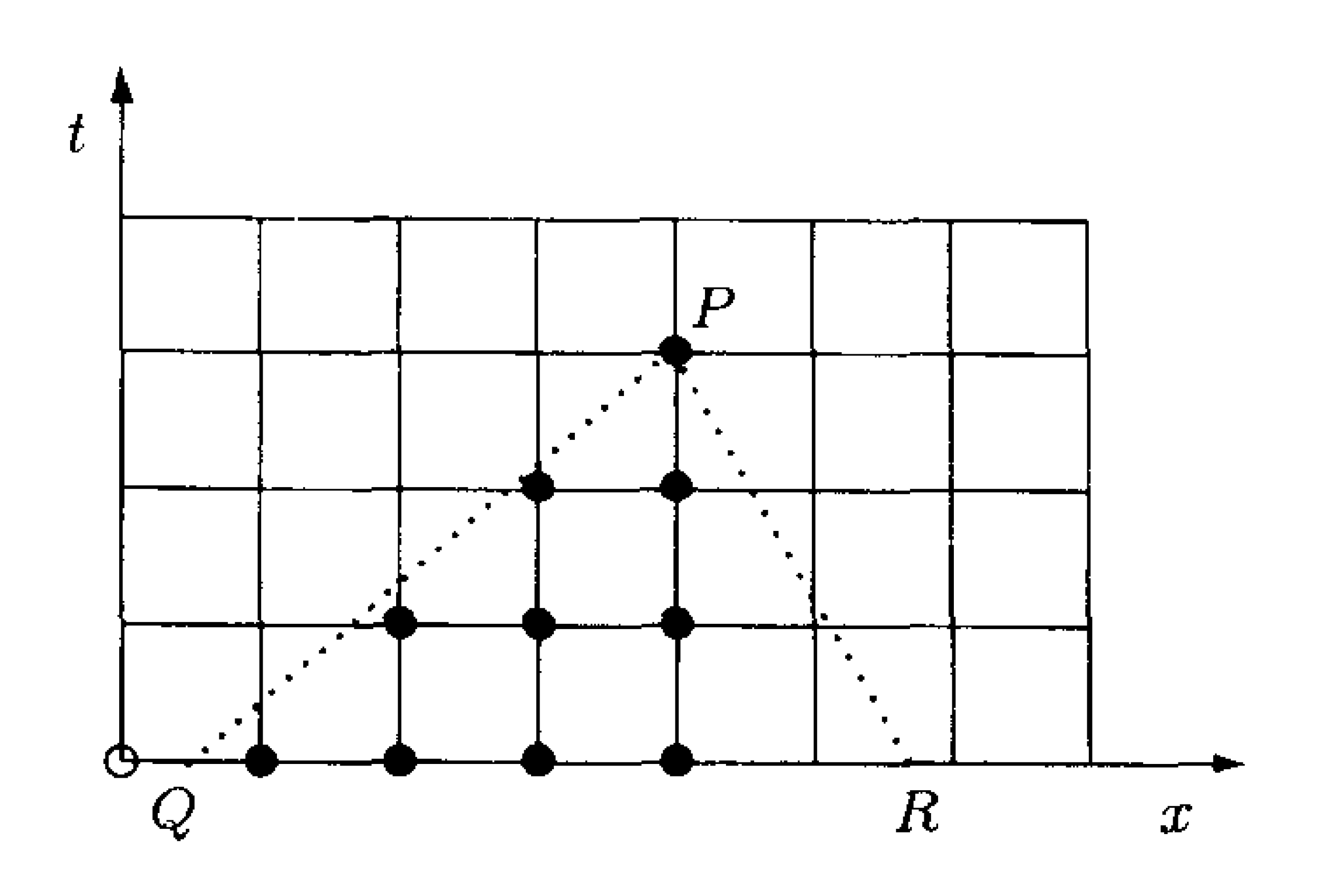


图 4-3　违背 CFL 条件的情形

通过这个例子的 CFL 条件，可以看出当 $a<0$ 时，该微分方程的特征线形如 RP，上述格式不可能收敛. 当 $a>0$ 时，特征线必须落在三角形依赖区域内，这就给出了时间步大小的一个约束，即 $a\Delta t/\Delta x\leqslant 1$.

这样得到的 CFL 条件可视为差分格式稳定性的一个必要条件，与第 2 章中抛物型方程显式格式的稳定性条件有点相似，而显然该条件对于变系数的方程甚至非线性方程都可以应用. 然而它仅是个 *必要*(necessary) 条件. 一般地，CFL 条件并不是稳定性的充分条件，通过以后的例子可看出这一点. CFL 条件的最大优点是其简单性，通过它只需稍稍考察就可否定很多差分格式. 而对满足 CFL 条件的格式，则还需用保证稳定性的某种检验法进一步考察其细节.

假定现在用一个较一般的显式格式近似对流方程 (4.1)，其中只用到上一时间层的三个对称分布的点，如图 4-4 所示；此时 CFL 条件为

$$|a|\Delta t\leqslant\Delta x, \tag{4.16}$$

常称 $\nu:=|a|\Delta t/\Delta x$ 为 *CFL数*(CFL number).

当 $a>0$ 时，为得到 U_j^{n+1} 必须用到 U_{j-1}^n 和 U_j^n; 当 $a<0$ 时，必须用到 U_j^n 和 U_{j+1}^n，

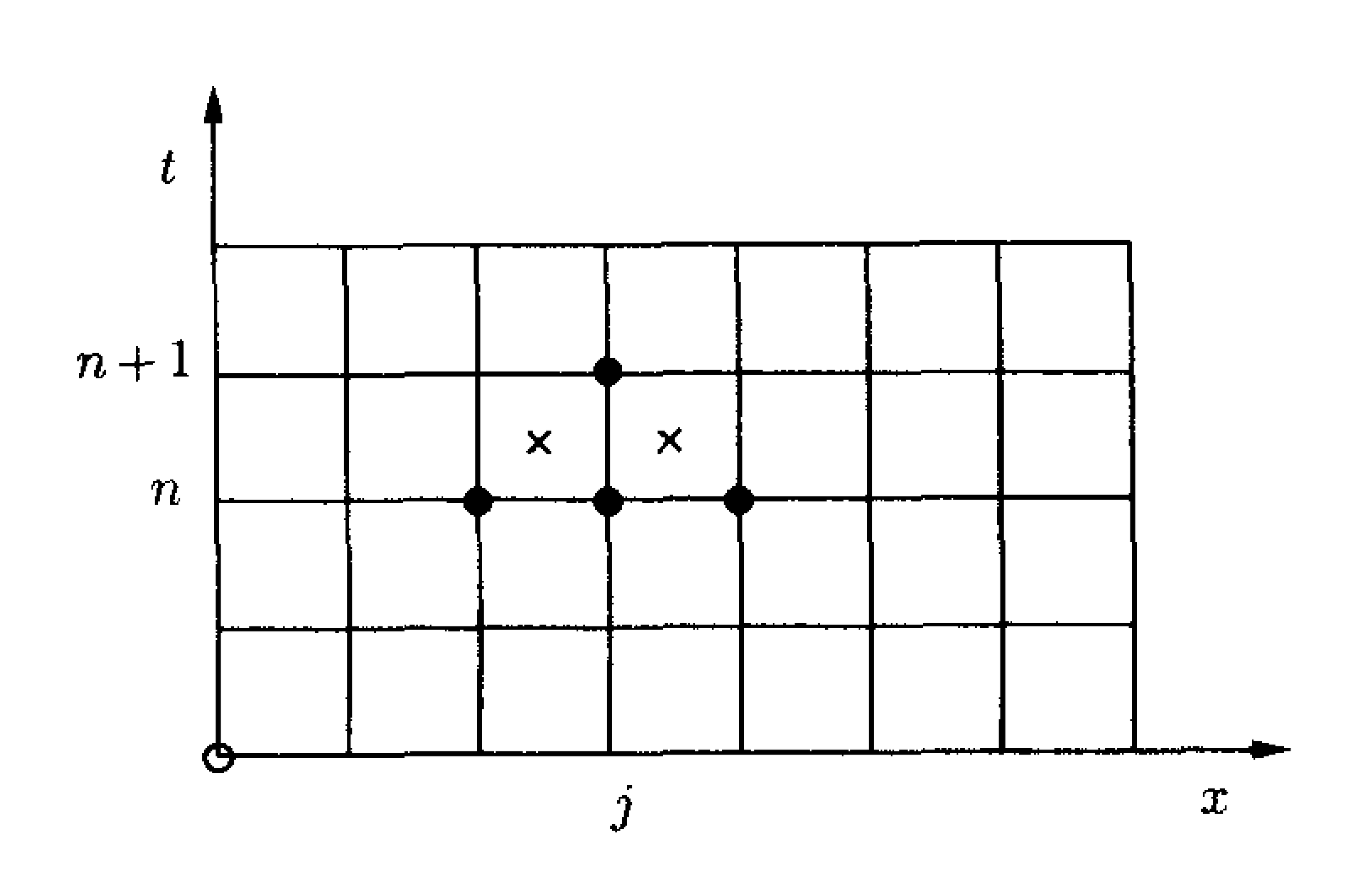


图 4-4　一般的三点格式；标记 × 的点用于两步 Lax-Wendroff 方法

为同时满足这两个要求，对空间导数用中心差分格式逼近，对时间导数用向前差分格式逼近，从而得到

$$\frac{U_j^{n+1}-U_j^n}{\Delta t}+a\frac{U_{j+1}^n-U_{j-1}^n}{2\Delta x}=0. \tag{4.17}$$

可看出若 (4.16) 式成立，则无论 a 取正还是取负，格式都满足 CFL 条件.

若 a 为常数，且不考虑边界条件的影响，则可像第 2 章中抛物型方程那样用傅里叶分析的方法来考察格式的稳定性. 傅里叶波型

$$U_j^n=(\lambda)^n \mathrm{e}^{\mathrm{i}\,k(j\Delta x)} \tag{4.18}$$

满足格式 (4.17) 的条件是，其增长因子 λ 满足

$$\lambda\equiv\lambda(k)=1-(a\Delta t/\Delta x)\mathrm{i}\,\sin k\Delta x. \tag{4.19}$$

因此，对任意的网格比 (对几乎所有的波型) 都有 $|\lambda|>1$，从而在保持网格比 $a\Delta t/\Delta x$ 不变的情况下，沿任何加密路径格式都是不稳定的. 注意到其中有一种情况是：对于最高频波型，此时 $k\Delta x=\pi$ 或 $U_j\propto(-1)^j$，振幅不增；但对于满足 $k\Delta x=\frac{1}{2}\pi$ 或 U_j 取一列值 $\ldots,-1,0,1,0,-1,\ldots$ 的波型，在向右移动的同时其模将每步增长 $[1+(a\Delta t/\Delta x)^2]^{1/2}$ 倍. 因此，虽然中心差分格式满足 CFL 条件，但它却不稳定. 这就证实了以前的说法，即 CFL 条件是稳定性的必要条件，但不充分.

只涉及到 3 个点的最简单而紧凑的稳定差分格式称为*迎风格式*(upwind scheme), 这是因为当 a 为正数时该格式在空间上用的是向后差分，当 a 为负数时用的是向前差分：

$$U_j^{n+1}=\begin{cases}U_j^n-a\dfrac{\Delta t}{\Delta x}\Delta_{+x}U_j^n, & 若\ a<0,\\[2ex] U_j^n-a\dfrac{\Delta t}{\Delta x}\Delta_{-x}U_j^n, & 若\ a>0.\end{cases} \tag{4.20}$$

若 a 不是常数，而是一个关于 x 和 t 的函数时，我们必须明确指定 (4.20) 式中 a 的取值. 当前不妨假定所取的值为 $a(x_j,t_n)$，在不产生混淆的情况下可舍去上下标仍记为 a, 并与 (4.15) 一样记 $\nu=a\Delta x/\Delta t$.

显然，当 (4.16) 成立时，迎风格式满足 CFL 条件，并且当常数 $a>0$ 时，由傅里叶分析可得增长因子为

$$\lambda\equiv\lambda(k)=1-(a\Delta t/\Delta x)(1-e^{-\mathrm{i}\,k\Delta x})\equiv 1-\nu(1-e^{-\mathrm{i}\,k\Delta x}). \tag{4.21}$$

从而有

$$\begin{aligned}|\lambda(k)|^2 &= [(1-\nu)+\nu\cos k\Delta x]^2+[\nu\sin k\Delta x]^2\\ &= (1-\nu)^2+\nu^2+2\nu(1-\nu)\cos k\Delta x\\ &= 1-2\nu(1-\nu)(1-\cos k\Delta x),\end{aligned}$$

即

$$|\lambda|^2 = 1 - 4\nu(1-\nu)\sin^2\frac{1}{2}k\Delta x. \tag{4.22}$$

所以当 $0 \leqslant \nu \leqslant 1$ 时，对所有的 k 都有 $|\lambda| \leqslant 1$. 当 $a < 0$ 时，同样的分析可得到几乎一样的增长因子 $\lambda(k)$ ，仅需将其中的 a 替换为 $|a|$. 因此，对迎风格式来说， CFL 条件给出了正确的稳定性界限，与 (2.56) 引入的冯诺伊曼条件相同.

4.3 迎风格式的误差分析

迎风格式 (4.20) 也可写为

$$U_j^{n+1} = \begin{cases} (1+\nu)U_j^n - \nu U_{j+1}^n, & \text{若 } a < 0, \\ (1-\nu)U_j^n + \nu U_{j-1}^n, & \text{若 } a > 0. \end{cases} \tag{4.23}$$

对之可作如下理解：如图 4-5 所示，当 $a > 0$ 时，设过点 $P = (x_j, t_{n+1})$ 的特征线与前一时间层的直线 $t = t_n$ 交于点 Q ，由 CFL 条件可知其必落在点 $A = (x_j, t_n)$ 和 $B = (x_{j-1}, t_n)$ 之间，又由于精确解 $u(x,t)$ 沿特征线是常数，故 $u(P) = u(Q)$. 若已知时间层 t_n 上所有节点的数值解，则可通过插值求出 $u(Q)$ ，并由此确定所求的值 U_j^{n+1} ，假定所用的是线性插值，即通过 A,B 点上的数值解构造一个关于 x 的线性函数来逼近 $u(x, t_n)$ ，当 a 为常数时， (4.23) 式的右端正是精确线性插值，因为此时 $AQ = \nu\Delta x$, $QB = (1-\nu)\Delta x$; 当 a 是光滑函数时，它也是一个好的近似.

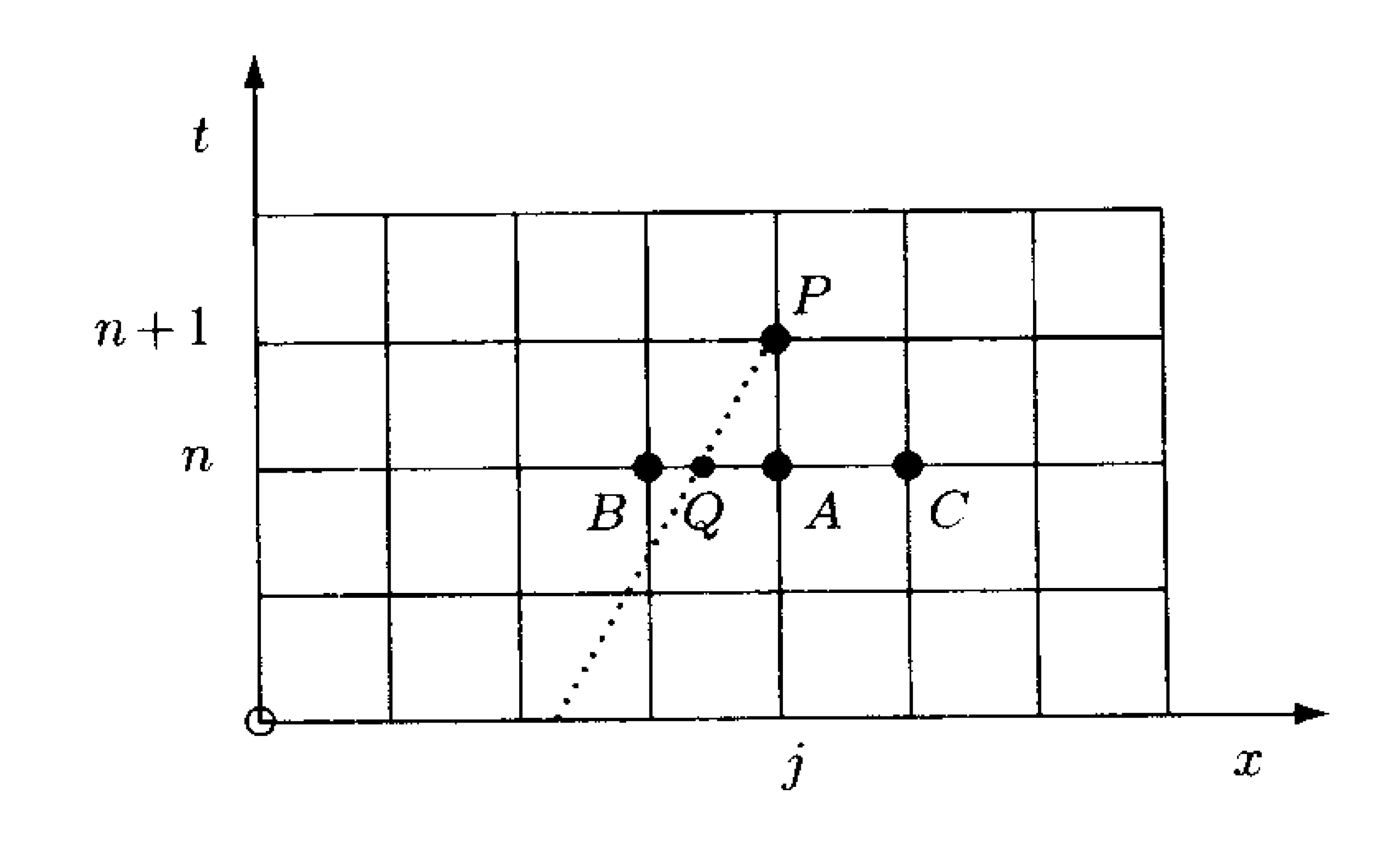


图 4-5 通过线性插值或二次插值构造差分格式

又注意到对于格式 (4.23) ，若在所有网格节点上都有 $|\nu| \leqslant 1$ ，则该格式中的系数均为非负，从而最大值原理成立，由此可以得到线性变系数问题的误差界，其方法与前面讨论抛物型方程时相同. 首先需仔细考查问题所在的区域，以及应给的边界条件，尽管有时物理问题定义在整条直线上，但对于所有的 x 值，数值解必须局限于有界区域中. 例如，假定感兴趣的区域是 $0 \leqslant x \leqslant X$ ，那么其边界为 $x = 0$ 和 $x = X$. 由于所考查的是一

阶双曲型微分方程，故通常只需一个边界条件．这与第 2 章的抛物型方程有着本质的区别，当时必须对区域的两端都给出边界条件．通过特征线的方向，可以看出：若在 $x=0$ 处有 $a>0$，则需给出 $x=0$ 上的边界条件；若在 $x=X$ 处有 $a<0$，则需给出 $x=X$ 上的边界条件；而当 a 在各处的符号一致时，则只需一个边界条件．这样，为得到微分方程在某点 P 的精确解，可从该点做特征线，该特征线与初始直线 $t=0$ 相交，或者与已给出边界条件的边界相交，从而解就由交点的值确定．

为简单起见，先假定在 $[0,X]\times[0,t_F]$ 上恒有 $a>0$，对于一般情形以后再考虑．按通常意义定义格式的截断误差，设 u 充分光滑，将其在 (x_j,t_n) 处展开得

$$\begin{aligned} T_j^n &:= \frac{u_j^{n+1}-u_j^n}{\Delta t}+a_j^n\frac{u_j^n-u_{j-1}^n}{\Delta x} \\ &\sim [u_t+\frac{1}{2}\Delta t\,u_{tt}+\cdots]_j^n+[a(u_x-\frac{1}{2}\Delta x\,u_{xx}+\cdots)]_j^n \\ &= \frac{1}{2}(\Delta t\,u_{tt}-a\,\Delta x\,u_{xx})+\cdots. \end{aligned} \tag{4.24}$$

即便 a 为常数，因而有 $u_{tt}=a^2u_{xx}$，我们仍然得到

$$T_j^n=-\frac{1}{2}(1-\nu)a\,\Delta x\,u_{xx}+\cdots;$$

因此，该格式一般只具有一阶精度．假定将该差分格式应用到点 $x_j=j\Delta x$ 上，其中 $j=1,2,\ldots,J$，且 $J\Delta x=X$，给定边界条件 $U_0^n=u(0,t_n)$，那么误差 $e_j^n=U_j^n-u_j^n$ 满足

$$e_j^{n+1}=(1-\nu)e_j^n+\nu e_{j-1}^n-\Delta tT_j^n, \tag{4.25}$$

且 $e_0^n=0$. 由此可以推出，若所有点上 $0\leqslant\nu\leqslant1$ 均成立，则有

$$E^{n+1}:=\max_j|e_j^{n+1}|\leqslant E^n+\Delta t\max_j|T_j^n|.$$

再假定截断误差有界，即对该区域中所有的 j 和 n 有

$$|T_j^n|\leqslant T, \tag{4.26}$$

则由通常的归纳方法可得，当 $U_j^0=u^0(x_j)$ 时，

$$E^n\leqslant n\Delta tT\leqslant t_FT. \tag{4.27}$$

由此即可证明，只要微分方程的解的二阶导数有界，则沿任何满足 CFL 条件的网格加密路径，迎风格式具有一阶收敛速度．

现在考虑一般情形，即在各节点上 a 的取值为一般数集

$$\{a_j^n:=a(x_j,t_n);\ j=0,1,\ldots,J\}.$$

显然在这些节点上成立一个与 (4.25) 类似的方程，若 $a_j^n\geqslant0$ 且 $j>0$，则 (4.25) 成立；若 $a_j^n\leqslant0$ 且 $j<J$，则把 (4.25) 中的 e_{j-1}^n 替换为 e_{j+1}^n 后所得相应的迎风方程成立；对于其

他情形，即 $a_0^n > 0$ 或 $a_J^n < 0$ 时，则由相应流入边界条件知， $e_0^{n+1} = 0$ 或 $e_J^{n+1} = 0$ 成立.之后的讨论可仿照上面的方法进行，则仍将得到相似的结果.

在第 2 章中曾提到，在对抛物型方程做分析时，若给定的条件中函数不连续或者其导数不连续，就会产生困难. 因为如果区域内部方程的解的导数并不连续，就难以估计此导数的界. 双曲型方程的情形却非常不同. 正如所知，模型问题的解沿特征线保持不变，假定初始函数 $u(x,0) = u^0(x)$ 的一阶导数在 $x = \xi$ 处具有跳跃间断，那么显然解 $u(x,t)$ 在过点 $(\xi, 0)$ 的特征线上也具有相似的间断，且该间断并不局限在区域边界上. 这样的处处满足 (4.4b) 的解，在间断线之外也处处满足微分方程；因此可以认为它定义了微分方程的*广义解* (generalised solution)，以区别于处处都满足微分方程的*经典解*(classical solution). 通过这种方法，确实可以对初始数据 $u^0(x)$ 本身具有间断的情形定义一个解.

大多数实际问题中出现的双曲型方程组，往往涉及不连续情形，或者初始数据不连续，或者解会产生新的不连续. 对于这种问题，上述的全局截断误差及最大模误差分析方法不再适用，这是由于分析中需用到的区域内解的导数并不处处存在. 因此截断误差只能用来做局部分析，且在任何情况下，通过做特征线，把 u_j^{n+1} 替换为特征线与 t_n 时间层交点上的值 $u(Q)$ 的方法，在做局部误差分析时都是比较可取的. 为分析某种方法的整体性质，或者把不同的方法进行比较，比较令人满意的做法有傅里叶分析，守恒性质分析或者修正方程分析等等 (可参见后面章节内容) .

4.4 迎风格式的傅里叶分析

由于双曲型方程常用来描述波的传播和发展，故傅里叶分析对于研究其数值方法的精度和稳定性有着十分重要的作用. 其中，增长因子 $\lambda(k)$ 的模描述了格式中波的*衰减* (damping)，而其辐角则描述了格式中的*色散* (dispersion)，即波传播的速度随其频率变化的程度. 为使当前的分析严格起见，须假定 a 是 (正) 常数. 傅里叶波型

$$u(x,t) = e^{\mathrm{i}(kx+\omega t)} \tag{4.28}$$

是微分方程 (4.1) 的精确解，只要 ω 和 k 满足*色散关系*(dispersion relation)

$$\omega = -ak. \tag{4.29}$$

该波型完全没有耗散，其振幅为常数；每时间步相位改变量是 $-ak\Delta t$. 与此相对应，若 (4.21) 成立，则相应的傅里叶波型 (4.18) 满足迎风格式，从而有 (4.22)，这表明，除了 $\nu = 1$ 的特别情况外，该波型都会有衰减. 数值波型的相位由下式给出

$$\arg\lambda = -\tan^{-1}\left[\frac{\nu\sin k\Delta x}{(1-\nu)+\nu\cos k\Delta x}\right]. \tag{4.30}$$

我们需要特别考虑 $k\Delta x$ 较小时的情形，因为正是这些波型可在网格上得到较好的逼近.

为了讨论迎风格式以及后面其他的格式，下面给出一个简单的引理：

引理 4.1. 若当 $p \to 0$ 时，q 能展开为 p 的幂级数形式

$$q \sim c_1 p + c_2 p^2 + c_3 p^3 + c_4 p^4 + \cdots,$$

那么有

$$\tan^{-1} q \sim c_1 p + c_2 p^2 + (c_3 - \frac{1}{3}c_1^3)p^3 + (c_4 - c_1^2 c_2)p^4 + \cdots.$$

其证明留作练习.

现将 (4.30) 展开，并应用上面引理，得到

$$\begin{aligned} \arg\lambda &\sim -\tan^{-1}[\nu(\xi - \frac{1}{6}\xi^3 + \cdots)(1 - \frac{1}{2}\nu\xi^2 + \cdots)^{-1}] \\ &= -\tan^{-1}[\nu\xi - \frac{1}{6}\nu(1-3\nu)\xi^3 + \cdots] \\ &= -\nu\xi[1 - \frac{1}{6}(1-\nu)(1-2\nu)\xi^2 + \cdots], \end{aligned} \tag{4.31}$$

其中记

$$\xi = k\Delta x. \tag{4.32}$$

$\nu = 1$ 的情形显然是相当特殊的，此时格式给出了精确解. 除此之外，迎风格式总具有*振幅误差* (amplitude error)，且由 (4.22) 式可知，该误差在每一时间步是 ξ^2 阶的，从而在整体上是 ξ 阶的；又由 (4.31) 式可知，迎风格式具有 ξ^2 阶的*相对相位误差* (relative phase error)，其符号依赖于 ν 的值，且当 $\nu = \frac{1}{2}$ 时，该项误差消失. 在第 4.11 节中将对振幅误差和相对相位误差做更详细的定义和更多的解释，并对不同格式做相应的比较.

图 4-6 展示了迎风格式的一些结果，考虑的问题是求解方程

$$u_t + a(x,t)u_x = 0, \quad x \geqslant 0,\ t \geqslant 0, \tag{4.33a}$$

其中

$$a(x,t) = \frac{1+x^2}{1+2xt+2x^2+x^4}, \tag{4.33b}$$

带有初始条件

$$u(x,0) = \begin{cases} 1, & \text{如果 } 0.2 \leqslant x \leqslant 0.4, \\ 0, & \text{否则}, \end{cases} \tag{4.34a}$$

和边界条件

$$u(0,t) = 0. \tag{4.34b}$$

上述问题的精确解是

$$u(x,t) = u(x^*, 0), \tag{4.35a}$$

其中

$$x^* = x - \frac{t}{1+x^2}. \tag{4.35b}$$

由于 $a(x,t) \leqslant 1$，计算时令 $\Delta t = \Delta x$，这显然满足 CFL 稳定性条件. 此方程的解表示一

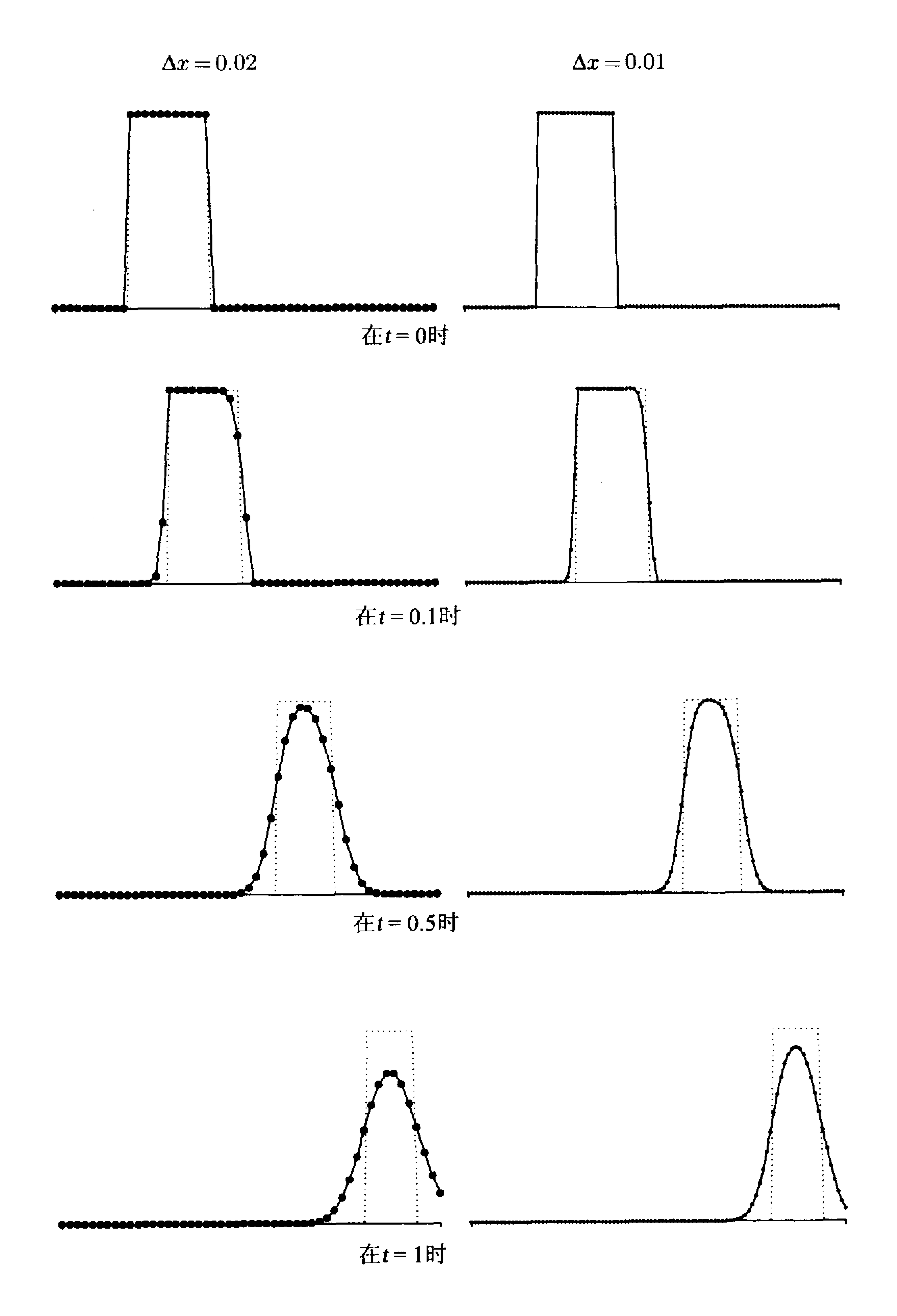


图 4-6 用迎风格式求解线性对流方程：问题 (4.33),(4.34)

向右移动的方形脉冲波．从图中显然可以看出，由于高频波的衰减而造成脉冲边缘有了实质性的磨光，且方波高度也略有下降．但是由于相位误差相当小，这使得脉冲按差不多正确的速度移动．图中后一部分是把时间和空间步长减半后所得的结果，虽然仍不太令人满意，但其精度较前者确实有了提高．

4.5　Lax-Wendroff 格式

实际上，迎风格式的相位误差比许多高阶格式的都小，但对大多数问题来说，因其造成的衰减过于严重而不可取．这时，我们可通过高阶插值得到更精确的显性格式，从图 4-5 中可以看出我们是如何使用线性插值近似 $u(Q)$ 而导出迎风格式的．仍假定特征线是斜率为 ν 的直线，现在用三点 A,B 和 C 上的值来作二次插值，可以得到 $u(Q)$ 的一个更精确的近似值．由此便得到 Lax-Wendroff 格式，它在本学科中有着极其重要的地位，最早在 1960 年，其创造者在研究双曲守恒律时就研究并应用了该格式 [1]．其形式为

$$U_j^{n+1} = \frac{1}{2}\nu(1+\nu)U_{j-1}^n + (1-\nu^2)U_j^n - \frac{1}{2}\nu(1-\nu)U_{j+1}^n, \tag{4.36}$$

亦可写为

$$U_j^{n+1} = U_j^n - \nu\Delta_{0x}U_j^n + \frac{1}{2}\nu^2\delta_x^2 U_j^n. \tag{4.37}$$

由通常的傅里叶分析知其增长因子为

$$\lambda(k) = 1 - \mathrm{i}\,\nu\sin k\Delta x - 2\nu^2\sin^2\frac{1}{2}k\Delta x. \tag{4.38}$$

分离实虚部，简单计算得

$$|\lambda|^2 = 1 - 4\nu^2(1-\nu^2)\sin^4\frac{1}{2}k\Delta x. \tag{4.39}$$

由此可知，当 $|\nu| \leqslant 1$ 时，即在满足 CFL 条件的整个范围内，格式是稳定的．还可得到

$$\begin{aligned} \arg\lambda &= -\tan^{-1}\left[\frac{\nu\sin k\Delta x}{1-2\nu^2\sin^2\frac{1}{2}k\Delta x}\right] \\ &\sim -\nu\xi[1-\frac{1}{6}(1-\nu^2)\xi^2+\cdots]. \end{aligned} \tag{4.40}$$

与迎风格式相比，该格式虽然仍有衰减，这是因为一般总有 $|\lambda| < 1$，但当 ξ 充分小时，其每步的振幅误差是 ξ^4 阶的，相对于迎风格式的 ξ^2 阶有了本质的提高．这两种格式的相对相位误差都是 ξ^2 阶的，且当 $\nu \sim 0$ 时二者是渐近相等的；但 Lax-Wendroff 的相对相位误差总不变号（对应于一种相位滞后效应），而对迎风格式来说，这种误差在 $\nu = \frac{1}{2}$

[1] Lax, P.D. and Wendroff, B.(1960), Systems of conservation laws, *Comm. Pure and Appl. Math.* **13**, 217-37.

处穿过零值. 尽管如此， Lax-Wendroff 格式衰减非常小所带来的好处，常常盖过了其相位误差较大的负面影响.

上面推导 Lax-Wendoff 格式时，假定 a 是常数. 为处理线性方程 (4.1) 中 a 是变量的情形，下面从原始的求导数的角度出发用另一种方法重新推导之. 首先关于变量 t 作泰勒展开，即

$$u(x,t+\Delta t)=u(x,t)+\Delta t\,u_t(x,t)+\frac{1}{2}(\Delta t)^2u_{tt}(x,t)+O((\Delta t)^3). \tag{4.41}$$

然后利用原微分方程把关于 t 的导数化为关于 x 的导数，即

$$u_t=-a\,u_x, \tag{4.42a}$$

$$u_{tt}=-a_tu_x-a\,u_{xt}, \tag{4.42b}$$

$$u_{xt}=u_{tx}=-(a\,u_x)_x, \tag{4.42c}$$

由此得

$$u_{tt}=-a_tu_x+a\,(a\,u_x)_x. \tag{4.43}$$

用中心差分格式逼近上式中所有关于 x 的导数，便得到格式

$$U_j^{n+1}=U_j^n-a_j^n\Delta t\frac{\Delta_{0x}U_j^n}{\Delta x}+\frac{1}{2}(\Delta t)^2\left[-(a_t)_j^n\frac{\Delta_{0x}U_j^n}{\Delta x}+a_j^n\frac{\delta_x(a_j^n\delta_xU_j^n)}{(\Delta x)^2}\right]. \tag{4.44}$$

该格式涉及到函数 $a(x,t)$ 在点 $x=x_j\pm\frac{1}{2}\Delta x$ 上的值以及点 x_j 上 a 和 a_t 的值，然而不难看出，若把 $\Delta_{0x}U_j^n$ 的系数 $a_j^n+\frac{1}{2}\Delta t(a_t)_j^n$ 用 $a_j^{n+1/2}$ 替代，格式可被简化；也可参见下面两节对守恒律 $au_x\equiv f_x$ 及有限体积法的讨论.

图 4-7 是用该格式求解问题 (4.33),(4.34) 所得的结果，前面为检验格式的性能，我们也曾用迎风格式在相同的网格尺度下求解过此问题. 比较图 4-6 和图 4-7 中的结果，可以发现在保持脉冲高度和宽度方面， Lax-Wendroff 格式比迎风格式效果要好得多，后者使得方波散开得很厉害. 另一方面，当脉冲向右传播时， Lax-Wendroff 格式算得的结果在两个间断后面产生了振荡. 也可看出减小网格步长 Δx ，虽然的确提高了计算结果的精度，但好像没有什么可以体现出格式误差阶 Δx^2 所预言的倍数 4. 这是因为截断误差分析只在解足够光滑的问题时有效，而该问题的解却有间断. 事实上对于该问题，迎风格式的最大误差是 $O((\Delta x)^{1/2})$ 阶的， Lax-Wendroff 格式最大误差是 $O((\Delta x)^{2/3})$ 阶的，因此网格步长减小时误差趋于零的速度相当缓慢.

图 4-7 中之所以产生振荡，是因为 Lax-Wendroff 格式不满足最大值原理. 从 (4.36) 可以看出，当 $\nu>0$ 时，由于稳定性又要求 $\nu\leqslant 1$ ，从而 U_{j+1}^n 的系数是负的. 也就是说，U_{j+1}^n 由上一时间层的三个值加权平均得到，但其中有两个加权因子为正，有一个为负. 因此数值解就可能产生振荡，即产生内部极大值和极小值.

作为有光滑解的问题的例子，考虑同样的问题 (4.33 a,b)，但把初始条件 (4.34) 改为

$$u(x,0)=\exp[-10(4x-1)^2]. \tag{4.45}$$

结果如图 4-8 所示，与前面类似，其解是一个向右传播的脉冲，但该脉冲具有光滑的高斯形状，而非不连续的方波形状．在同样的网格尺寸下，结果要精确得多．到了 $t=1$ 时刻，脉冲左端仍出现了一点振荡，但比起不连续情形时要小得多．而且加密网格确实减小了误差，振荡也几乎消失了．

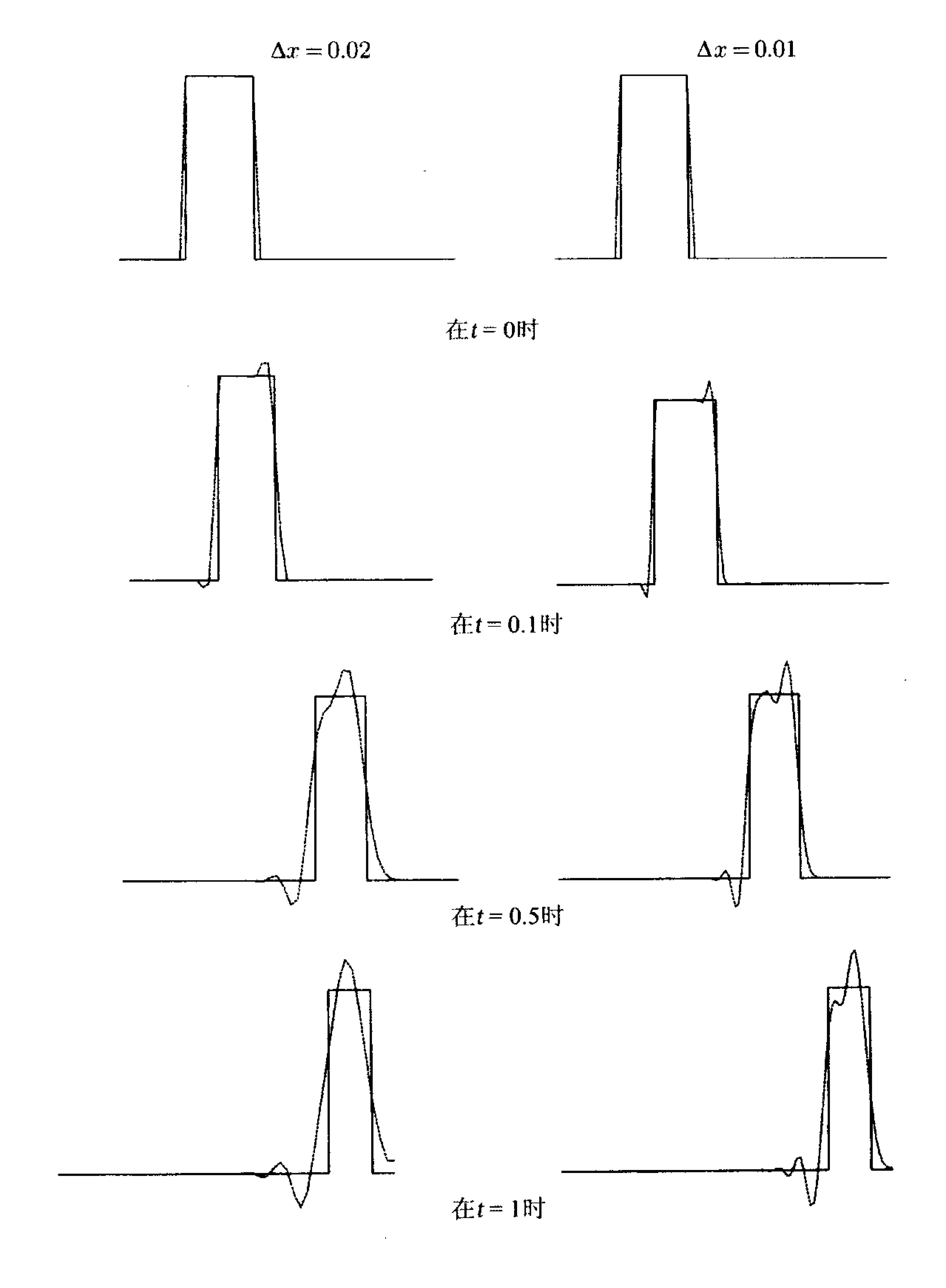


图 4-7　Lax-Wendroff 方法求解线性对流方程：问题 (4.33),(4.34)

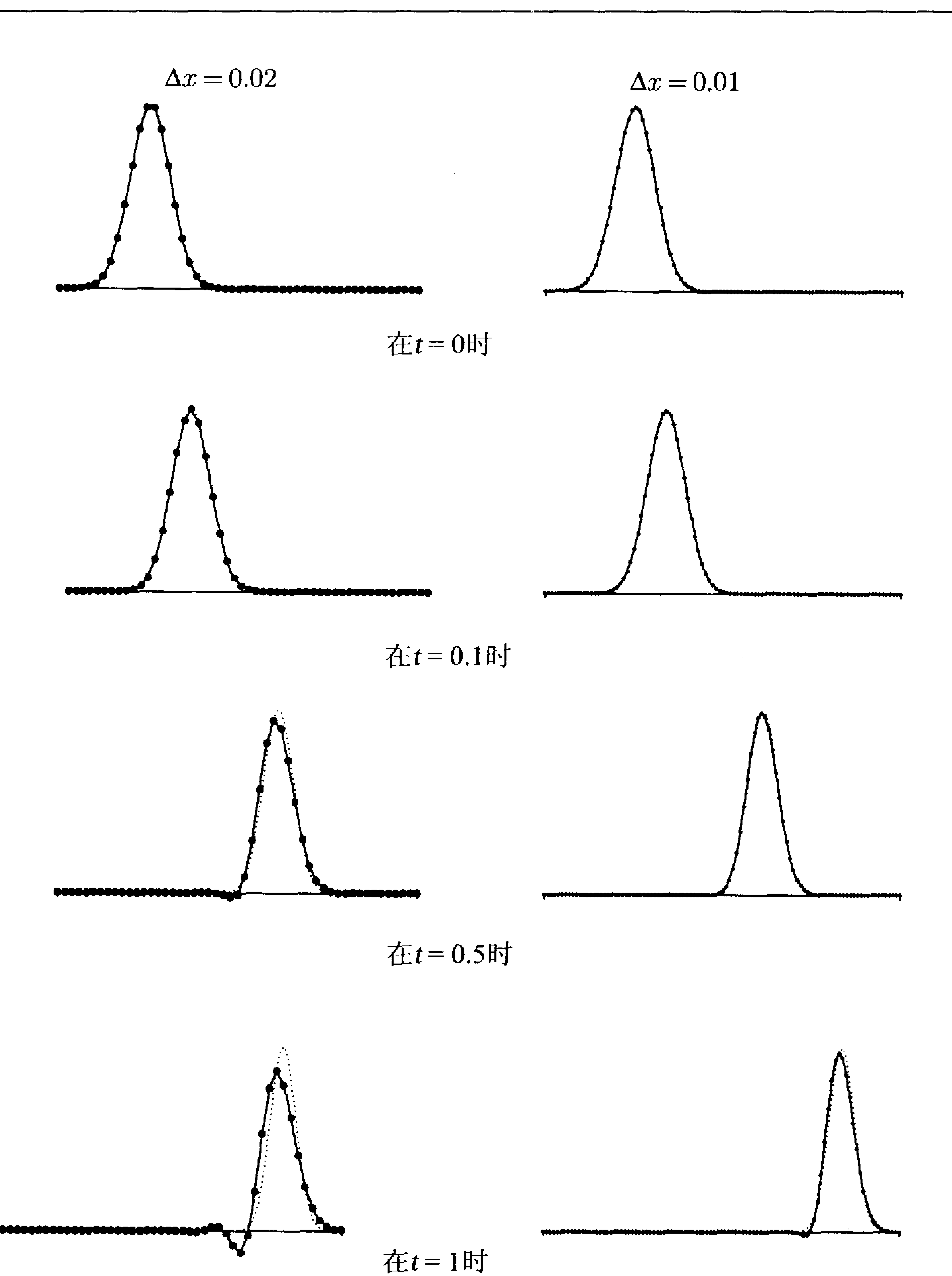


图 4-8 Lax-Wendroff 方法求解线性对流方程：问题 (4.33), 初始条件 (4.45)

4.6 守恒律的 Lax-Wendroff 方法

在实际应用中，双曲型方程常以如下形式出现

$$\frac{\partial u}{\partial t}+\frac{\partial f(u)}{\partial x}=0. \tag{4.46}$$

上式也可写成以前所考察的形式

$$u_t + au_x = 0, \tag{4.47}$$

其中 $a = a(u) = \partial f/\partial u$. 直接从 (4.46) 推导 Lax-Wendroff 格式是很方便的. 显然, 函数 f 不依赖于 x 和 t , 仅是 u 的函数. 泰勒展开式 (4.41) 中所需的关于 t 的导数可写为

$$u_t = -(f(u))_x \tag{4.48a}$$

以及

$$u_{tt} = -f_{xt} = -f_{tx} = -(au_t)_x = (af_x)_x. \tag{4.48b}$$

如前所做, 用中心差分替代关于 x 的导数, 得

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}\Delta_{0x} f(U_j^n) + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 \delta_x[a(U_j^n)\delta_x f(U_j^n)]. \tag{4.49}$$

显然当 a 是常数, 且 $f(u) = a\,u$ 时, 上式可化简为 (4.37) . 把 (4.49) 的最后一项展开, 可以发现其中包括 $a(U_{j-1/2}^n)$ 和 $a(U_{j+1/2}^n)$ 的值, 为求之, 可令 $U_{j\pm1/2}^n := \frac{1}{2}(U_j^n + U_{j\pm1}^n)$, 但更常用的方法是用 $\Delta_{\pm x} f(U_j^n)/\Delta_{\pm x} U_j^n$ 替代. 记 F_j^n 表示 $f(U_j^n)$, $A_{j\pm1/2}^n$ 分别表示上述两种特征速度, 则格式化为

$$U_j^{n+1} = U_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left\{\left[1 - A_{j+1/2}^n \frac{\Delta t}{\Delta x}\right]\Delta_{+x} F_j^n + \left[1 + A_{j-1/2}^n \frac{\Delta t}{\Delta x}\right]\Delta_{-x} F_j^n\right\}. \tag{4.50}$$

作为此格式应用的一个例子, 考虑 Burgers' 方程的极限情形, 对于无粘流体, 有

$$u_t + u\, u_x = 0, \tag{4.51}$$

或写成守恒形式

$$u_t + (\frac{1}{2}u^2)_x = 0. \tag{4.52}$$

当广义解光滑时, 则很容易由特征线方法求之, 或者只需验证其解可由下面的隐式给出

$$u \equiv u(x,t) = u^0(x - t\,u(x,t)). \tag{4.53}$$

其特征线为直线, 其解沿着每条特征线为常数. 事实上, 给定初始条件 $u(x,0) = u^0(x)$, 对任意的 x_0 , 过点 $(x_0, 0)$ 以 $\mathrm{d}t/\mathrm{d}x = 1/u^0(x_0)$ 为斜率做直线, 该直线即为特征线.

假设给定光滑初始条件 (4.45), 我们就会发现对于非线性双曲型方程常出现的一种典型的困难. 由于解沿每条特征线都为常数, 所以当特征线相交时就会出现奇性. 事实上, 既然特征线相交, 其在相交点上必有不同的斜率, 于是解沿着这些特征线必有不同的值, 这样, 解在交点处就是多值的. 当初始条件 $u^0(x)$ 的导函数递减时, 这种现象总要出现. 图 4-9 画出了一些典型的特征线, 对此类问题总存在一个临界值 t_c , 使得当 $0 \leqslant t < t_c$ 时, 解存在且为单值函数, 而在 $t = t_c$ 时刻开始出现奇性.

这种现象是气流中激波(shock) 形成的一个简单模型. 当特征线相交时, 不仅 (4.51) 的经典解本身不存在, 而且数学模型对物理现象的解释也已失效. 当解在空间的变化梯度很大时, 粘性将起到重要的作用, 此时必须考虑带粘性的完整的 Burgers' 方程 $u_t+uu_x=\nu_v u_{xx}$. 对此种情况的完整的讨论已超出本书的范围, 但有一些关键点与我们强调用方程的守恒律形式是非常相关的.

在超过临界点之后, 我们希望得到的是粘性方程 ν_v 趋于零时解的极限. 此时将产生一个间断, 代表一个激波, 对于守恒律 $u_t+f_x=0$, 其以*激波速度* (shock speed)

$$S:=\frac{[f(u)]}{[u]} \tag{4.54}$$

移动, 其中 $[u]$ 表示变量 u 的跳跃. 因此若 u 在间断左边的极限值是 u_L , 在右端的极限值是 u_R , 那么激波将以速度

$$\frac{f(u_R)-f(u_L)}{u_R-u_L}$$

移动. 当 $u_R\to u_L$ 时, 即激波减 "弱" 时, 激波速度显然趋于 $a(u_L)$. 上述激波速度关系式可由以下过程推导. 在 (x,t) 平面内沿特征线取一块盒式区域, 并使之覆盖一部分特征线, 在该区域中对原方程积分, 再应用高斯散度定理, 就可得到所要的结果. 这种仅在平均意义下满足微分方程的解称作*弱解* (weak solution).

然而, 应该注意到这样得到的速度依赖于守恒律的具体形式, 例如从非守恒型形式 (4.51) 出发, 可以导出以下的任何一种守恒律形式

$$((m+1)u^m)_t+(mu^{m+1})_x=0. \tag{4.55}$$

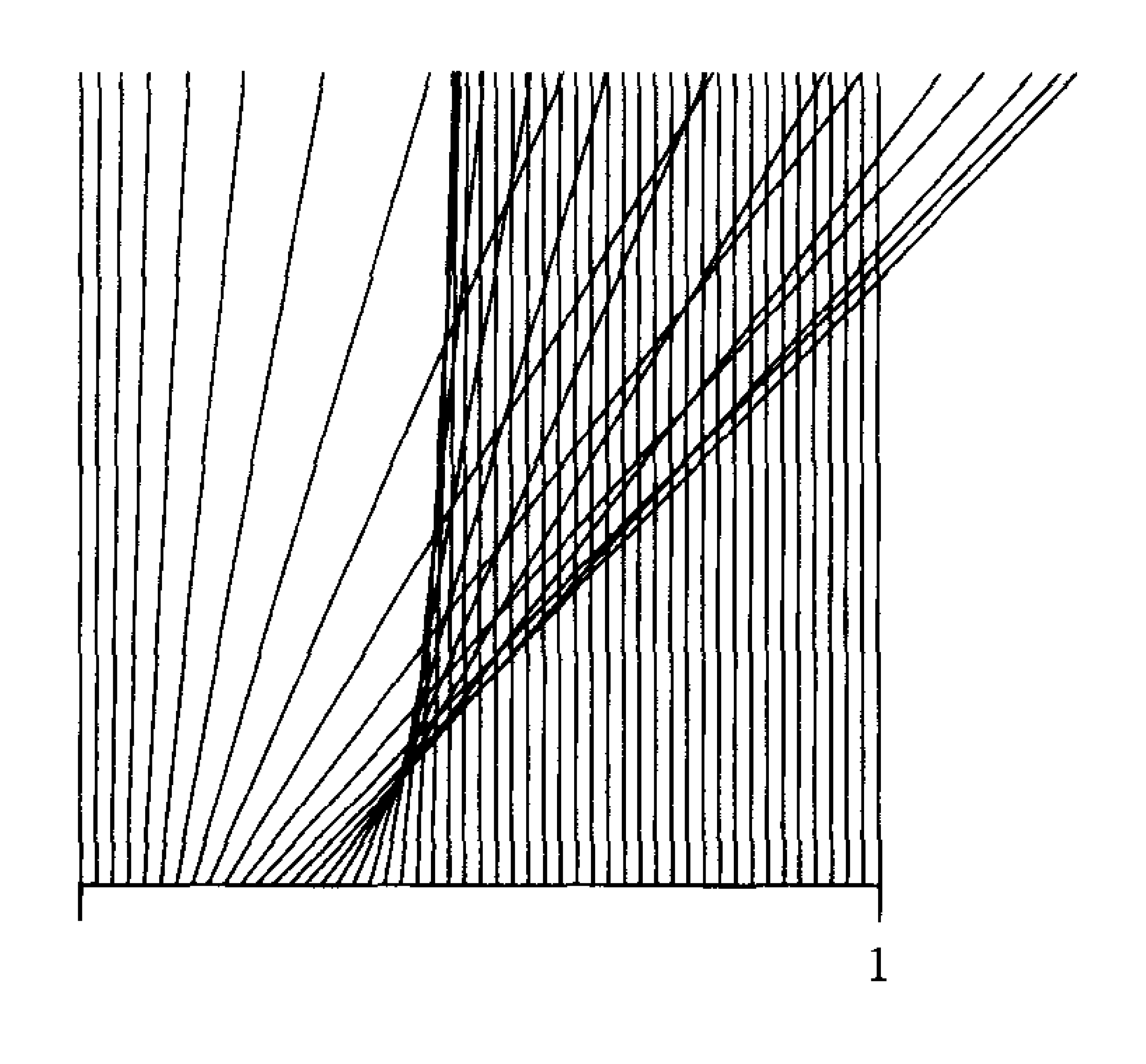


图 4-9 无粘 Burgers 方程的典型的特征线

对于相当复杂的物理现象，我们必须选择正确的模型. 对于当前情形，$m = 1$ 即 (4.52) 是正确的，它并不等价于 (4.51)，而是比之更具一般性. 由此知，从 u_L 到 u_R 的激波将以平均速度 $\frac{1}{2}(u_L+u_R)$ 传播，而不是其他速度，例如不是 $m = 2$ 时的速度 $\frac{2}{3}(u_L^2+u_Lu_R+u_R^2)/(u_L+u_R)$.

图 4-10 展示了计算守恒律 (4.52) 所得的刚开始的一段时间内的结果，其初始条件为

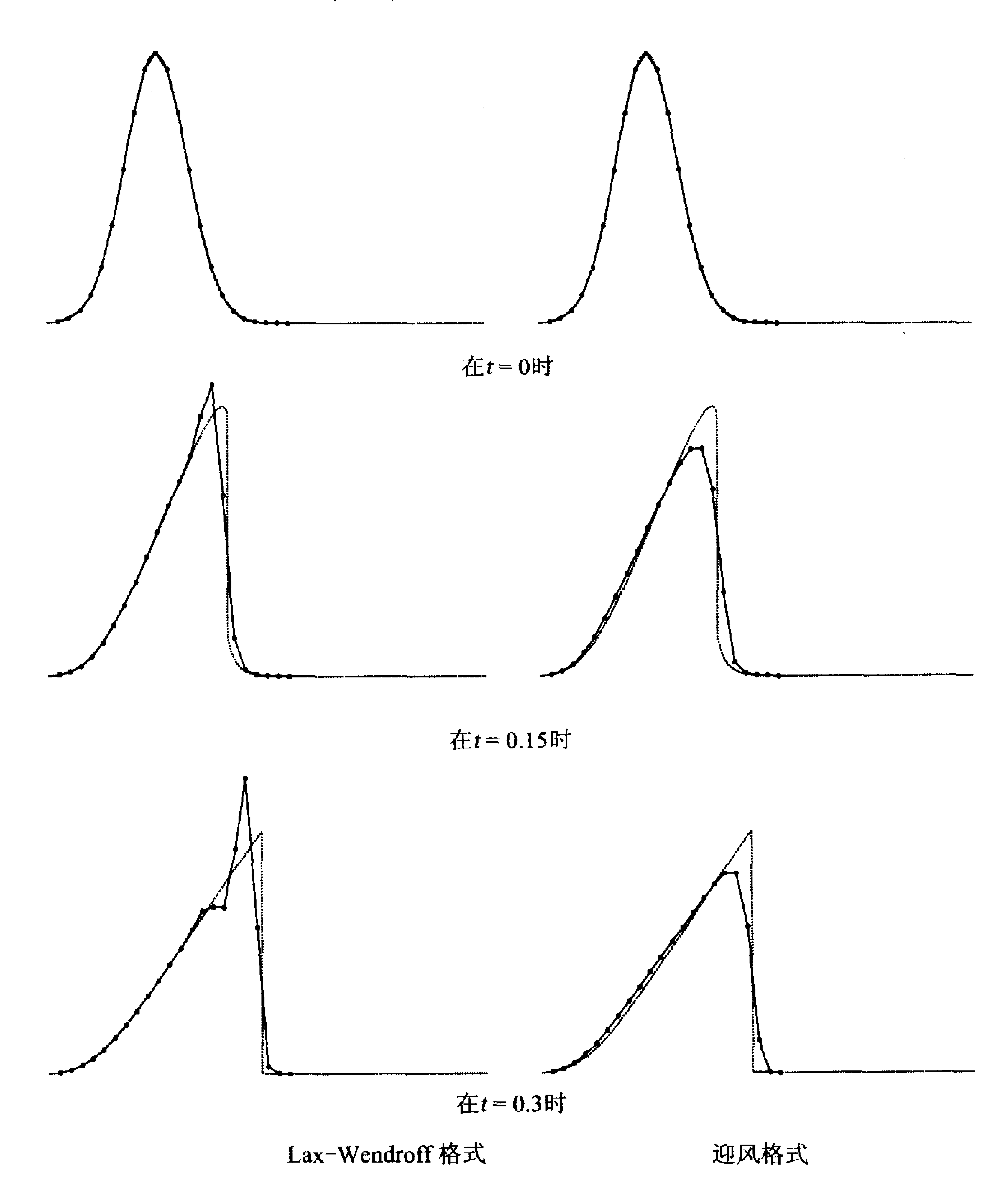


图 4-10　用 Lax-Wendroff 格式(左)和迎风格式(右)求解带初始条件 (4.45) 的 Burgers' 方程

(4.45). 大约在 $t_c = 0.25$ 时开始出现激波，其强度也在增长，但是解在整体上将最终衰减为零. 方程的弱解如虚线所示，左边显示的是用 Lax-Wendroff 格式逼近所得的结果. 为

了比较，在图 4-10 右边给出了用迎风格式逼近所得的结果，这里迎风格式的形式如下

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{1}{2}\frac{\Delta t}{\Delta x}\left\{\left[1-\operatorname{sgn} A_{j+\frac{1}{2}}^n\right]\Delta_{+x}F_j^n\right. \\ &\qquad \left. + \left[1+\operatorname{sgn} A_{j-\frac{1}{2}}^n\right]\Delta_{-x}F_j^n\right\}. \end{aligned} \tag{4.56}$$

其中较好的选择是取 $A_{j\pm\frac{1}{2}}^n := \Delta_{\pm x}F_j^n/\Delta_{\pm x}U_j^n$，这当 $U_j^n = U_{j\pm1}^n$ 时简化为 $a(U_j^n)$. 此格式显然是 (4.20) 式的推广，且可以与 (4.50) 式作直接比较. 从图中显然可以看出 Lax-Wendroff 格式在远离激波时有更高的精度，但在激波后面却出现振荡，而迎风格式却没有出现振荡. 这就促使产生了一种在计算过程中自动选择使用这两种格式的思想，这种思想由 van Leer[1] 的工作所首创.

Lax-Wendroff 格式的一大优点是可以很容易地将之推广到方程组的情形. 代替 (4.46) 和 (4.48 a,b), 有

$$\mathbf{u}_t = -\mathbf{f}_x,\ \mathbf{u}_{tt} = -\mathbf{f}_{tx} = -(A\mathbf{u}_t)_x = (A\mathbf{f}_x)_x, \tag{4.57}$$

其中 A 是 (4.8) 所示的雅可比矩阵， (4.49) 式则可简单地改写为

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \left(\frac{\Delta t}{\Delta x}\right)\Delta_{0x}\mathbf{f}(\mathbf{U}_j^n) + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2\delta_x\left[A(\mathbf{U}_j^n)\delta_x\mathbf{f}(\mathbf{U}_j^n)\right], \tag{4.58}$$

(4.50) 式几乎不用做任何改变，只是这里要使用的是向量 $\mathbf{U}_j^n$ 和 $\mathbf{F}_j^n$.

特别地，若 $\mathbf{f}(\mathbf{u}) = A\mathbf{u}$，其中 A 为常数矩阵，则格式简化为

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \left(\frac{\Delta t}{\Delta x}\right)A\Delta_{0x}\mathbf{U}_j^n + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 A^2\delta_x^2\mathbf{U}_j^n. \tag{4.59}$$

对该问题可做傅里叶分析. 向量 U 的每个分量都是同一个傅里叶波型的倍数，并寻找方程 (4.59) 的如下形式的解

$$\mathbf{U}_j^n = \lambda^n \mathrm{e}^{\mathrm{i}\,kj\Delta x}\widehat{\mathbf{U}} \tag{4.60}$$

其中 $\widehat{\mathbf{U}}$ 为一常数向量. 上式是解的条件为

$$\left\{\lambda I - \left[I - \mathrm{i}\left(\frac{\Delta t}{\Delta x}\right)\sin k\Delta x A - 2\left(\frac{\Delta t}{\Delta x}\right)^2\sin^2\frac{1}{2}k\Delta x A^2\right]\right\}\widehat{\mathbf{U}} = \mathbf{0}. \tag{4.61}$$

为使此式成立，$\widehat{\mathbf{U}}$ 必须是 A 的特征向量，设 μ 为相应的特征值，并记 $\nu = \mu\Delta t/\Delta x$，可得

$$\lambda = 1 - \mathrm{i}\,\nu\sin k\Delta x - 2\nu^2\sin^2\frac{1}{2}k\Delta x. \tag{4.62}$$

---

[1] van Leer, B.(1974), Towards the ultimate conservative difference scheme. II monotonicity and conservation combined in a second order scheme, *J. of Comput. Phys.* **14**, 361-70.

此式完全与 (4.38) 相同．由此可以推出该格式稳定的一个必要条件

$$\frac{\rho\Delta t}{\Delta x} \leqslant 1, \tag{4.63}$$

其中 ρ 表示 A 的谱半径 (spectral radius)，即 A 的所有特征值之模的最大值．这里我们直接将原来的稳定性概念推广到了方程组情形．而对冯诺伊曼条件在方程组情形是否也是稳定性的充分条件这个问题，将放到下章讨论．

Lax-Wendroff 格式的一个重要的优点是，其稳定性条件只涉及到特征值的模，而不依赖于其正负符号，因此格式的形式并不用随着正负符号的改变而改变．从 (4.20) 和 (4.56) 式可知，对于单个方程，迎风格式必须根据 a 的符号选择用向前差分还是向后差分；对于方程组来说，这种选择将变得相当困难．因为此时我们必须在每一点都寻找一个近似雅可比矩阵 $\tilde{A}$ 的特征值和特征向量，并把当前向量表示成特征向量的线性组合，再根据相应特征值的符号决定每个特征向量用向前差分还是向后差分，最后在把这些特征向量组合起来得到新的时间层上的解． Lax-Wendroff 格式则避免了这种麻烦，但为此付出的代价是数值解如图 4-7 和图 4-10 所示出现了振荡．正是这种情况促使前面提到的结合这两种格式优点的方法得到了很大的进展，既然已经提到这一点，下节我们将对此做进一步的讨论．虽然大大超出了本书的范围，现在引入激波关系式 (4.54) 的推广仍然是值得注意的 (这种推广也是迎风格式 (4.56) 的特征速度 $A^n_{j\pm1/2}$ 所满足的关系式的推广)，即近似雅可比矩阵 $\tilde{A}$ 通常应该满足

$$\tilde{A}^n_{j\pm1/2}\Delta_{\pm x}\mathbf{U}^n_j = \Delta_{\pm x}\mathbf{F}^n_j. \tag{4.64}$$

上式是根据 Roe [1] 关于空气动力学的工作提出的，是开发守恒律的新方法时常考虑的一个重要的关系式．

最后， Lax-Wendroff 格式的一个很方便且常用的形式是两步格式，第一步给出点 $(x_{j+1/2}, t_{n+1/2})$ 上的值，第二步利用中心差分最终得到 t_{n+1} 时刻的值（参见图 4-4）：

$$\mathbf{U}^{n+1/2}_{j+1/2} = \frac{1}{2}(\mathbf{U}^n_j + \mathbf{U}^n_{j+1}) - \frac{1}{2}(\Delta t/\Delta x)\left[\mathbf{f}(\mathbf{U}^n_{j+1}) - \mathbf{f}(\mathbf{U}^n_j)\right], \tag{4.65a}$$

$$\mathbf{U}^{n+1}_j = \mathbf{U}^n_j - (\Delta t/\Delta x)\left[\mathbf{f}(\mathbf{U}^{n+1/2}_{j+1/2}) - \mathbf{f}(\mathbf{U}^{n+1/2}_{j-1/2})\right]. \tag{4.65b}$$

对线性情形 $\mathbf{f} = A\mathbf{u}$ ，其中 A 是常数矩阵，从上面两步格式中消去中间变量 $\mathbf{U}^{n+1/2}_{j+1/2}$，就可得到标准的单步 Lax-Wendroff 格式 (4.59) ，其证明留作练习．对于非线性或者变系数的情形，这两种形式并不等价． (4.65) 的最重要的优点是它避免了计算雅可比矩阵 A ．

[1] Roe, P.L.(1981), Approximate Riemann Solvers,parameter vectors,and difference schemes, *J. of Comput. Phys.* **43**, 357-72.

4.7 有限体积格式

许多用于守恒律实际计算的方法都可归类为有限体积方法(finite volume method),(4.65)就是一个典型的例子. 假定给出守恒律方程组的守恒形式 $\mathbf{u}_t+\mathbf{f}_x=0$，将之在 (x,t) 平面中的一个区域 Ω 中积分，并利用高斯散度定理将之化为线积分

$$\begin{aligned}\int\int_\Omega(\mathbf{u}_t+\mathbf{f}_x)\,\mathrm{d}x\,\mathrm{d}t &\equiv \int\int_\Omega \mathrm{div}(\mathbf{f},\mathbf{u})\,\mathrm{d}x\,\mathrm{d}t\\ &= \oint_{\partial\Omega}[\mathbf{f}\,\mathrm{d}t-\mathbf{u}\,\mathrm{d}x].\end{aligned} \tag{4.66}$$

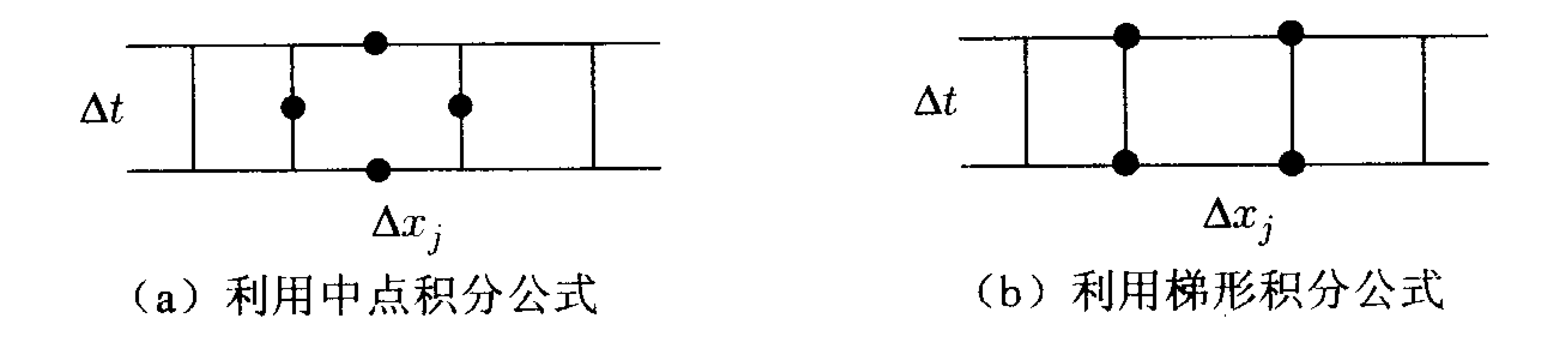
特别地，如所选的区域是宽为 Δx, 高为 Δt 的矩形区域，再引进各边上的每个变量的平均值，诸如 $\mathbf{u}_{上}$ 等等，于是得到

$$(\mathbf{u}_{上}-\mathbf{u}_{下})\Delta x+(\mathbf{f}_{右}-\mathbf{f}_{左})\Delta t=0. \tag{4.67}$$

为了得到具体的数值格式，这些平均值需要用某种形式的积分公式代替. 例如，可以在每条边上用中点积分公式 (如图 4.11(a)), 设 $\mathbf{U}_j^n$ 表示第 n 时间层上宽度为 Δx_j 的第 j 个单元的中点上的近似值， $\mathbf{F}_{j+1/2}^{n+1/2}$ 表示通量在该单元右侧边中点上的值，于是得到格式

$$\mathbf{U}_j^{n+1}=\mathbf{U}_j^n-(\Delta t/\Delta x_j)\left(\mathbf{F}_{j+1/2}^{n+1/2}-\mathbf{F}_{j-1/2}^{n+1/2}\right). \tag{4.68}$$

下面所剩的是用若干 $\mathbf{U}_j^n$ 的值来计算通量，例如可利用两步 Lax-Wendroff 格式中所用的泰勒展开式计算：即先用公式 (4.65a) 计算单元侧边上的解，再将之带入到方程 (4.65b) 中，便得到了形如 (4.68) 的公式.



(a) 利用中点积分公式　　(b) 利用梯形积分公式

图 4-11　有限体积格式

不过，应该注意到，在 (4.68) 中单元宽度是完全任意的，这也是该格式的一个很大的优点，在实际计算中非常有用 (特别是在高维情形) . 例如，从 (4.68) 出发，把一串相邻单元的积分加起来，便得到

$$\sum_{j=k}^{l}\Delta x_j(\mathbf{U}_j^{n+1}-\mathbf{U}_j^n)+\Delta t(\mathbf{F}_{l+1/2}^{n+1/2}-\mathbf{F}_{k-1/2}^{n+1/2})=0. \tag{4.69}$$

此式精确反映了微分方程的守恒性质. 为推导 Lax-Wendroff 格式，设 $\mathbf{U}_j^n$ 表示单元中心

的值，在单元侧边 $x_{j+1/2}$ 上把解做一阶精度的泰勒展开

$$\mathbf{u}(x_{j+1/2}, t_n + \Delta t/2) = \mathbf{u}(x_{j+1/2}, t_n) - \frac{1}{2}\Delta t\, \mathbf{f}_x(x_{j+1/2}, t_n) + O((\Delta t)^2);$$

再与各单元中心上的解在各边的泰勒展开式联立，就可得到公式

$$\mathbf{U}_{j+1/2}^{n+1/2} = \frac{\Delta x_{j+1}\mathbf{U}_j^n + \Delta x_j \mathbf{U}_{j+1}^n - \Delta t\left[\mathbf{f}(\mathbf{U}_{j+1}^n) - \mathbf{f}(\mathbf{U}_j^n)\right]}{\Delta x_j + \Delta x_{j+1}}. \tag{4.70}$$

该公式把 (4.65a) 推广到了一般网格上．为避免不必要的复杂性，下面仍回来考虑通常的均匀网格的情形．

如前所示， Lax-Wendroff 格式的主要缺点是其解容易产生振荡．这个问题促使了有限体积法的大力发展，在标量守恒律情形可以对该问题做完全的分析．其主要原理是要控制解的*总变差* (total variation)：把有限区域 $[0, X]$ 剖分为 J 个单元， $\mathbf{U}^n$ 在时间层 n 的第 j 个单元取值为 U_j^n ，由此定义解的总变差为

$$\mathrm{TV}(U^n) := \sum_{j=1}^{J-1} |U_{j+1}^n - U_j^n| \equiv \sum_{j=1}^{J-1} |\Delta_{+x} U_j^n|. \tag{4.71}$$

更一般地，对于精确解 $u(x,t)$,$\mathrm{TV}(u(\cdot,t))$ 可按如下方式定义：对 $[0, X]$ 所有的剖分，诸如 $0 = \xi_0 < \xi_1 < \cdots < \xi_K = X$ ，把相应的差分绝对值 $|u(\xi_{j+1}, t) - u(\xi_j, t)|$ 求和，再对所有这些和求上确界；把这个上确界称为精确解的总变差．显然当把 $\mathbf{U}^n$ 看作是 $u(\cdot, t_n)$ 的分段常数近似时，这两个定义是相容的．然而，为简化以后的讨论，先忽略具体的边界条件，而假设 $u(\cdot, t)$ 以及 $\mathbf{U}^n$ 都以常数向左右扩展，所以这里并未指定求和指标 j 的具体范围．

守恒律 (4.46) 的解的一个关键性质是，其总变差 $\mathrm{TV}(u(\cdot,t))$ 是 t 的非增函数 (这大致可从解沿特征线保持不变 (4.4b) 这个事实推出)．因此我们把满足 $\mathrm{TV}(U^{n+1}) \leqslant \mathrm{TV}(U^n)$ 的格式定义为 TVD*总变差不增*(total variation diminishing) 格式．这一概念最早由 Harten[1] 给出，他还建立了以下的一个有用的定理．

定理 4.1. **(Harten)** *一个差分格式是* **TVD** *的，如果它可写成如下形式*

$$U_j^{n+1} = U_j^n - C_{j-1}\Delta_{-x}U_j^n + D_j \Delta_{+x} U_j^n, \tag{4.72}$$

其系数 C_j 和 D_j 可以是 $\{U_j^n\}$ 的任意函数，满足

$$C_j \geqslant 0,\ D_j \geqslant 0 \text{ 且 } C_j + D_j \leqslant 1\ \ \forall j. \tag{4.73}$$

证明. 由向前差分格式 (4.72) ，利用恒等式

[1] Harten, A.(1984), On a class of high resolution total-variation-stable finite-difference schemes, *SIAM J. Numer. Anal.* **21**, 1-23.

$\Delta_{+x}U_j \equiv \Delta_{-x}U_{j+1}$ ，得到

$$
\begin{aligned}
U_{j+1}^{n+1} - U_j^{n+1} &= \Delta_{+x}U_j^n - C_j\Delta_{+x}U_j^n + C_{j-1}\Delta_{-x}U_j^n + D_{j+1}\Delta_{+x}U_{j+1}^n - D_j\Delta_{+x}U_j^n \\
&= (1 - C_j - D_j)\Delta_{+x}U_j^n + C_{j-1}\Delta_{-x}U_j^n + D_{j+1}\Delta_{+x}U_{j+1}^n.
\end{aligned}
$$

由条件 (4.73) 知，上面最后一个式子右端各项的所有系数都非负．于是，取绝对值可得

$$
|\Delta_{+x}U_j^{n+1}| \leqslant (1 - C_j - D_j)|\Delta_{+x}U_j^n| + C_{j-1}|\Delta_{-x}U_j^n| + D_{j+1}|\Delta_{+x}U_{j+1}^n|,
$$

再对 j 求和，一些项相互抵消之后得到 $\mathrm{TV}(U^{n+1}) \leqslant \mathrm{TV}(U^n)$. ■

现把该定理应用到 Lax-Wendroff 方法和迎风方法中．先考虑后者，迎风格式由 (4.56) 给出，其中 $A_{j\pm1/2}^n := \Delta_{\pm x}F_j^n/\Delta_{\pm x}U_j^n$, 此格式即是 Roe 提出的格式的标量情形，关于 Roe, 我们曾在 (4.64) 后面的几句话中提到．我们最好把此格式看作有限体积格式，在此格式中 (4.68) 的通量为

$$
F_{j+1/2}^{n+1/2} = \begin{cases} f(U_j^n), & \text{当 } A_{j+1/2}^n \geqslant 0 \text{ 时}, \\ f(U_{j+1}^n), & \text{当 } A_{j+1/2}^n < 0 \text{ 时}; \end{cases} \tag{4.74}
$$

或者等价地

$$
F_{j+1/2}^{n+1/2} = \frac{1}{2}\left[\left(1 + \mathrm{sgn}A_{j+1/2}^n\right)F_j^n + \left(1 - \mathrm{sgn}A_{j+1/2}^n\right)F_{j+1}^n\right]. \tag{4.75}
$$

然后，在 (4.56) 中用 $A_{j-1/2}^n\Delta_{-x}U_j^n$ 替代通量差 $\Delta_{-x}F_j^n$ ，再与 (4.72) 做比较，便得到

$$
C_{j-1} = \frac{1}{2}\frac{\Delta t}{\Delta x}\left(1 + \mathrm{sgn}A_{j-1/2}^n\right)A_{j-1/2}^n.
$$

它显然总非负的，因此满足 (4.73) 的第一个条件，类似地可以得到

$$
D_j = -\frac{1}{2}\frac{\Delta t}{\Delta x}\left(1 - \mathrm{sgn}A_{j+1/2}^n\right)A_{j+1/2}^n,
$$

它也是非负的．另外，将两式求和，并注意先将前者下标移位，即有 [1]

$$
\begin{aligned}
C_j + D_j &= \frac{1}{2}\frac{\Delta t}{\Delta x}\left[\left(1 + \mathrm{sgn}A_{j+1/2}^n\right)A_{j+1/2}^n + \left(1 - \mathrm{sgn}A_{j+1/2}^n\right)A_{j+1/2}^n\right] \\
&\equiv |A_{j+1/2}^n|\frac{\Delta t}{\Delta x},
\end{aligned}
$$

这恰恰是 CFL 数．因此， (4.73) 的最后一个条件对应于 CFL 稳定性条件；我们前面已经证明，当 Δt 满足稳定性条件时， Roe 一阶迎风差分格式是 TVD 的．

另一方面，对 Lax-Wendroff 格式 (4.50) 重复上一过程，记 $\nu_{j\pm1/2}^n$ 表示 $A_{j\pm1/2}^n\Delta t/\Delta x$ ，则可得到

$$
C_j = \frac{1}{2}\nu_{j+1/2}^n(1 + \nu_{j+1/2}^n), \quad \text{及} \quad D_j = -\frac{1}{2}\nu_{j+1/2}^n(1 - \nu_{j+1/2}^n), \tag{4.76}
$$

[1] 原书下式有误，写成了： $(1 + \mathrm{sgn}A_{j+1/2}^n)A_{j+1/2}^n + (1 - \mathrm{sgn}A_{j+1/2}^n)A_{j+1/2}^n$.

两者必须都是非负的， (4.73) 的第三个条件要求 CFL 条件 $(\nu_{j+1/2}^n)^2 \leqslant 1$ 成立，为同时满足这三个条件，则 $\nu_{j+1/2}^n$ 只能取 $-1, 0$ 或 $+1$ ，除了很特殊的情况外这显然不太实际.

Roe 迎风格式具有 TVD 性质,这使它成为发展更复杂的有限体积方法的重要基石. 特别是在计算激波时,它非常成功,然而当处理某些稀疏波时还必须对该格式做些改进. 例如,假设考虑无粘 Burgers 方程 (4.52) ，给定初始值 $\{U_j^0 = -1,$ 对 $j \leqslant 0;\ U_j^0 = +1,$ 对 $j > 0\}$ ，则其解为一逐渐展开的稀疏波. 然而，从 (4.74) 式显然可看到， Roe 格式的数值通量都为 $\frac{1}{2}$ ，因此其解相对于初值不会有任何改变. 产生这种问题与*声速点*(sonic point) 有关. 这里的声速点是指 $u=0$ 时的点，此时特征速度 a 也为零，更准确地说，对于*跨声速稀疏波*(transonic rarefaction wave) ，特征速度在该点左侧为负，右侧为正. 对于一般的凸通量函数 $f(u)$ ，假定 $u = u_s$ 是（唯一的）声速点，则需对有限体积法作如下改进，把其中的通量 (4.75) 改为

$$F_{j+1/2}^{n+1/2} = \frac{1}{2}\left[\left(1+\operatorname{sgn} A_j^n\right) F_j^n + \left(\operatorname{sgn} A_{j+1}^n - \operatorname{sgn} A_j^n\right) f(u_s) + (1-\operatorname{sgn} A_{j+1}^n) F_{j+1}^n\right]. \tag{4.77}$$

这种格式使用的是特征速度 $\{A_j^n = a(U_j^n)\}$ 的正负符号，而不是差商 $\{A_{j+1/2}^n\}$ 的正负符号，这种改进后的格式最早由 Engquist 和 Osher[1] 提出，并已被广泛研究和应用. 不难看出，只有当 U_j^n 和 U_{j+1}^n 之间存在声速点时，此格式才与 Roe 迎风格式有差别，它也是 TVD 的 (见习题 4.11) ，且能正确地计算跨声速稀疏波.

然而，这两种格式都是一阶精度的，事实上很难设计出二阶精度的 TVD 格式. 为理解其原因，不妨来考虑形如 (4.72) 的三点显式 TVD 格式，它满足条件 (4.73) . 对于线性对流方程 $u_t + au_x = 0$ ，假定 C 和 D 均为常数. 按照推导 Lax-Wendroff 格式 (4.36) 的方法，不难看出，若格式为二阶精度的，必然导致如 (4.76) 的系数，因此，除了很特别的情形外，其总违背 TVD 条件. 从另一个角度来看，在上述两种成功的 TVD 格式中，我们仅仅是用单元平均值 U_j^n 来计算数值通量的，而仅用分段常数近似是不可能得到二阶精度的.

这一视角指出了解决这种问题的办法，即引进被称为*重构* (recovery) 或*重建* (reconstruction) 的中间步，从 $\{U_j^n\}$ 出发构造 $u(\cdot, t_n)$ 的高阶逼近 $\tilde{U}^n(\cdot)$. 其中最著名的方法大概是 van Leer[2] 构造 MUSCL(Monotone Upstream-centred Schemes for Conservation Laws) 格

[1] Engquest, B. and Osher, O.(1981), One-sided difference approximations for non-linear conservation laws, *Math. Comp.* **36**, 321-52.

[2] van Leer, B.(1979), Towards the ultimate conservative difference scheme V. A second order sequel to Godunov's method, *J. of Comput. Phys.* **32**, 101-36.

式时所用的方法， 该方法用不连续的分段线性函数得到二阶逼近. 另一个精心设计的方法是 Colella 和 Woodward [1] 建立 PPM(Piecewise Parabolic Method) 格式时所用的，能达到三阶精度. 这几种方法中，重构都保证重构得到的函数的单元平均与原来相等. MUSCL 格式所用的方法如下：先找到横坐标为某单元中点、纵坐标为对应的单元平均值的点，再过该点作一条直线，其斜率要通过左右单元上函数平均值来确定. 对于 PPM 格式，先通过单元交点两侧的单元平均值确定该点上函数的近似值，再由端点上得到的值以及单元平均值在每个单元上确定一条抛物线，从而得到一个总体上连续的逼近函数. 对这些格式来说，重构后就仅需要将有限体积格式 (4.68) 中的数值通量作相应的改变. 如何利用重构逼近 $\tilde{U}^n(\cdot)$ 完成最后一步超出了本书的讨论范围，我们仅仅指出该过程一般基于求解一个局部发展问题，其具体解法源于 Godunov [2] 的研究工作. 但是，应注意到要如此得到 TVD 格式也必须对重构过程有所约束. 一个经典的约束为，重构应是*单调保持的* (monotonicity preserving)，也就是说，若 $\{U_j^n\}$ 单调增的，则 $\tilde{U}^n(\cdot)$ 也必须单调增.

4.8 盒式格式

为使读者对实际中使用的格式大概有所了解，此后给出两种非常重要的格式. 本节给出的*盒式格式*(box scheme) 是非常重要的一种隐式格式，常与 Thomée [3] 或 Keller [4] 的名字联系在一起. 对于最简单的模型问题 $u_t + au_x = 0$，其中 a 为常数，盒式格式形式如下

$$\frac{\delta_t\left(U_j^{n+1/2} + U_{j+1}^{n+1/2}\right)}{2\Delta t} + \frac{a\,\delta_x\left(U_{j+1/2}^n + U_{j+1/2}^{n+1}\right)}{2\Delta x} = 0. \tag{4.78}$$

引进平均算子

$$\mu_x U_{j+1/2} := \frac{1}{2}(U_j + U_{j+1}), \tag{4.79}$$

以及相似的算子 μ_t，可把该格式写成非常简洁的形式

$$(\mu_x\delta_t + \nu\mu_t\delta_x)\,U_{j+1/2}^{n+1/2} = 0, \tag{4.80}$$

[1] Colella, P. and Woodward, P.R.(1984), The piecewise parabolic method (PPM) for gas-dynamical simulations, *J. of Comput. Phys.* **54**, 174-201.

[2] Godunov, S.K.(1959), A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics, *Mat. Sb.* **47**, 271-306.

[3] Thomée, V.(1962), A stable difference scheme for the mixed boundary value problem for a hyperbolic first order system in two dimensions, *J. Soc. Indust. Appl. Math.* **10**,229-45.

[4] Keller, H.B.(1971), A new finite difference scheme for parabolic problems, in B. Hubbard(ed.), *Numerical Solution of Partial Differential Equations II, SYNSPADE 1970*, Academic Press,pp.327-50.

其中 $\nu = a\Delta t/\Delta x$ 是 CFL 数.

假若对格式中所有的项都在中心点 $(x_{j+1/2}, t_{n+1/2})$ 像以前一样做泰勒展开，由差分平均的对称性不难看出，所得到的展开式中将只含 Δx 和 Δt 的偶次幂项，从而格式具有二阶精度. 当系数 a 是 x 和 t 的函数时，在 (4.78) 中最好取 $a_{j+1/2}^{n+1/2} := a(x_{j+1/2}, t_{n+1/2})$，这样得到的泰勒展开式不变，格式仍然是二阶精度的.

对具有守恒形式 (4.46) 的非线性问题，相应的盒式格式可写为

$$\frac{\delta_t\left(U_j^{n+1/2} + U_{j+1}^{n+1/2}\right)}{2\Delta t} + \frac{\delta_x\left(F_{j+1/2}^{n} + F_{j+1/2}^{n+1}\right)}{2\Delta x} = 0, \tag{4.81}$$

其中 $F_j^n := f(U_j^n)$. 该格式显然是一个有限体积格式，对应的区域是以四个相邻网格节点为顶点的正方形，在各边上积分时使用的是梯形规则 (参见图 4-12 以及 4-11(b))；另外，和其他有限体积方法一样，该格式可被推广到非一致分布的网格上.

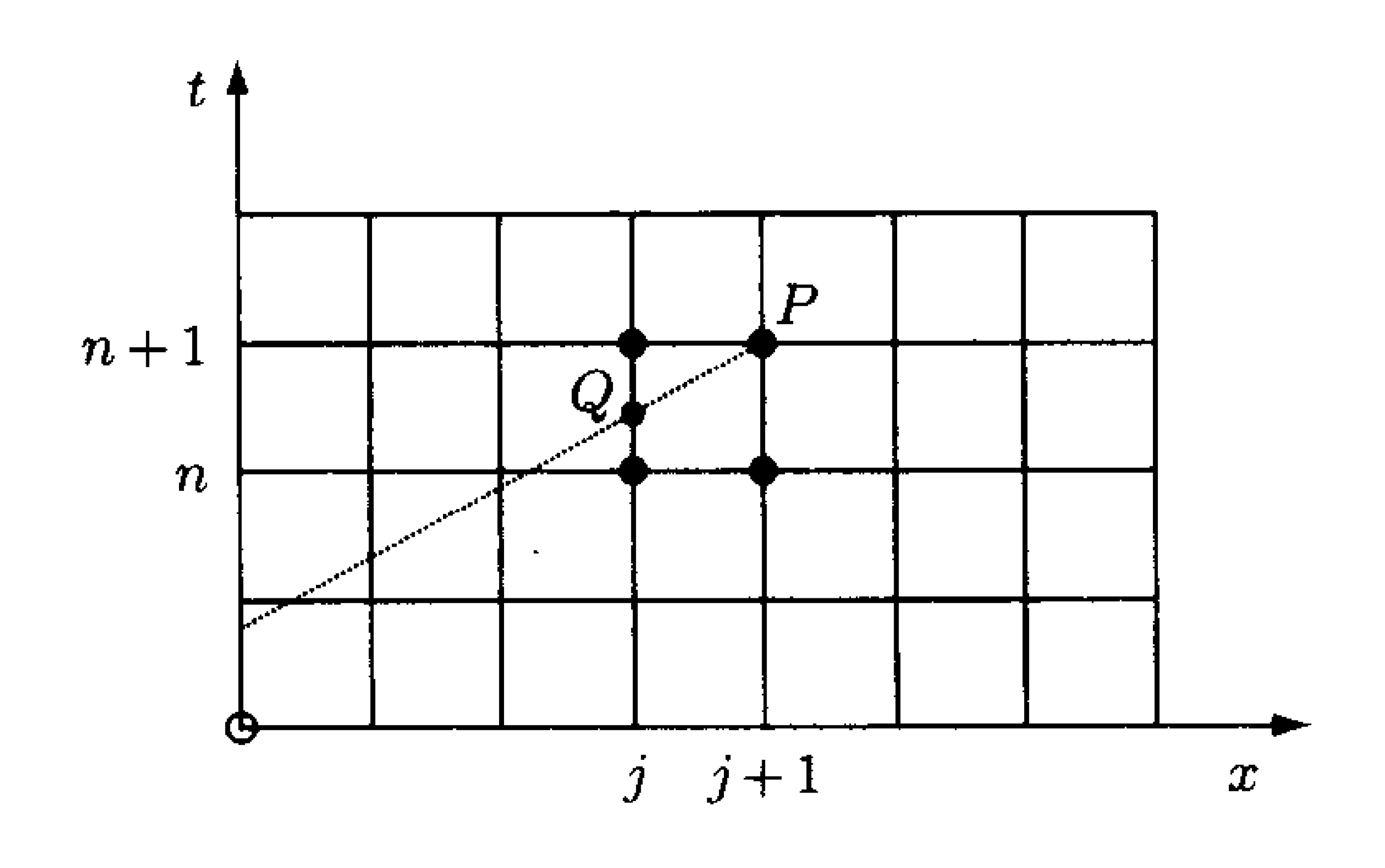


图 4-12　盒式格式

因为格式涉及到新时间层上的两个点，所以它是 隐式(implicit) 格式. 但对于最简单的模型问题，该格式并不需要额外的计算，并且只要使用方式正确，它还是无条件稳定的. (4.78) 可写为如下形式

$$U_{j+1}^{n+1} = U_j^n + \left(1 + \nu_{j+1/2}^{n+1/2}\right)^{-1}\left(1 - \nu_{j+1/2}^{n+1/2}\right)\left(U_{j+1}^n - U_j^{n+1}\right), \tag{4.82}$$

其中

$$\nu_{j+1/2}^{n+1/2} = a_{j+1/2}^{n+1/2}\frac{\Delta t}{\Delta x}.$$

当 $a(x,t)$ 为正时，特征速度也为正，此时必然已给出区域左端边界上的边界条件，也就是说，已给出新时间层上 U 的第一个值 U_0^{n+1}，通过 (4.82) 式，即可从左到右依次得到 U 的值. 当特征速度为负时，此时已给出右端边界条件，类似地可从右到左求解. 当方程为非线性守恒形式时，此格式却不太好用，这是由于为得到 U_{j+1}^{n+1} 的值，必须求解非线性方程 (4.81).

使用盒式格式的一个很严重的问题是,解有可能产生形如 $(-1)^{j+n}$ 的棋盘波型(chequerboard mode) 噪声污染. 这是因为格式 (4.78) 对时间和空间都取了平均, 该波型是方程的伪解波型, 只有利用边界条件和初始条件才能控制它的出现. 如图 4-13 所示, 对于方波初值和光滑初值, 前者解的噪声更为明显.

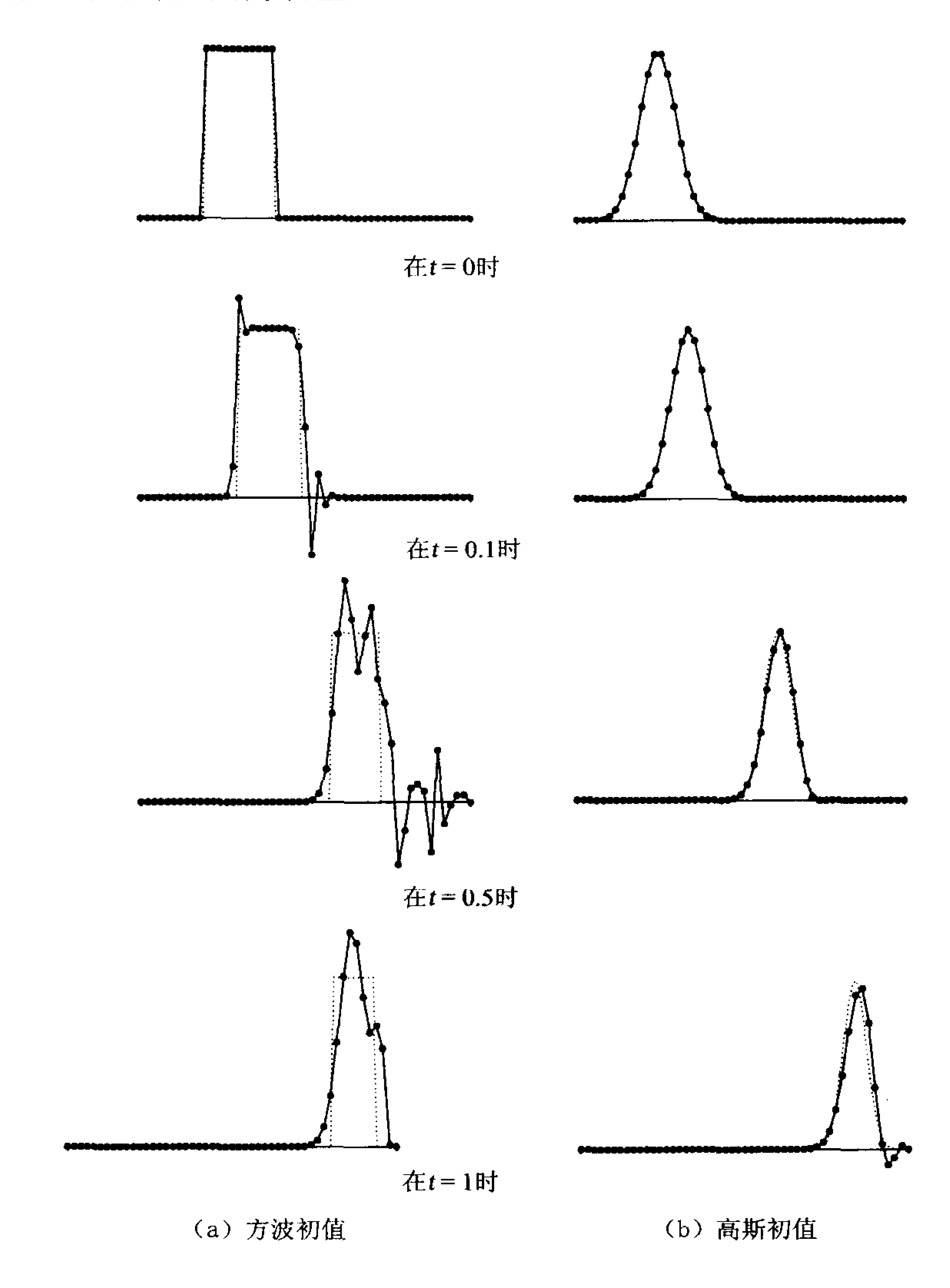


图 4-13 盒式格式求解线性对流方程, 其中 $\Delta t = \Delta x = 0.02$

对于形如 (4.5) 的方程组来说, 用该格式计算将变得很复杂, 因为此时格式是真正意义上的隐式格式. 一般地, 由于通常在 x 区域的两端都有边界条件, 已不太可能像标量形式时那样简单地在某个方向上扫描一遍就能得到解, 此时必须联立求解一组相应的方程才能得到新时间层上的解. 这和第 2 章热传导方程的隐式格式相似, 但却不那么直接.

方程组的矩阵一般是块三对角的，每一块的维数等于微分方程组的阶，但其具体的结构细节还依赖于在两端分别所给的边界条件的个数.

对于标量情形，只要我们按正确的方向使用此格式，例如若 a 为正，则使用 (4.82) 式，此时对任意的网格比 $\Delta t/\Delta x$， CFL 条件总是满足的. 如图 4-12 所示，特征线总落在构造 U_{j+1}^{n+1} 所用的三个点之间. 但是，该格式的三个系数未必都是正的，所以它不满足最大值原理；也没有任何自然的 TVD 性质，而是易于产生振荡解. 由于这里的区域既不是整条直线，也不满足周期性边界条件，所以直接应用傅里叶分析是不严格的，即便 a 是常数时也是如此. 即便如此，我们还是可以把某傅里叶波型代入，以考查其衰减程度以及相的精度. 不难看出

$$\lambda(k)=\frac{\cos\frac{1}{2}k\Delta x-\mathrm{i}\,\nu\sin\frac{1}{2}k\Delta x}{\cos\frac{1}{2}k\Delta x+\mathrm{i}\,\nu\sin\frac{1}{2}k\Delta x}, \tag{4.83}$$

从而可推出对任何的 ν, 都有

$$|\lambda(k)|=1, \tag{4.84}$$

并且

$$\begin{aligned}\arg\lambda &= -2\tan^{-1}\left(\nu\tan\frac{1}{2}k\Delta x\right)\\ &\sim -\nu\xi\left[1+\frac{1}{12}(1-\nu^2)\xi^2+\cdots\right].\end{aligned} \tag{4.85}$$

所以，盒式格式没有衰减波型，并且通过比较 (4.85) 和 (4.40) 可以看出，该格式的相同样具有二阶精度，但这里的相位误差大致是 Lax-Wendroff 格式的一半，另外也可看出，当 $\nu\leqslant 1$ 时，后者的相误差是滞后的，而盒式格式的则是提前的；而当 $\nu>1$ 时，盒式格式的相误差也变为滞后的了. 在第 4.11 节中可看到相误差的比较图. 在这里也应该注意到，因为盒式格式是无条件稳定的，故最好令 $|\nu|$ 接近于 1 ，这样能得到较高阶的相精度.

对于这样的无耗散 (nondissipative) 格式，我们可引入一个启发性概念 *群速度* (group velocity)，用于表示色散 PDE 中能量的传播速度. 若作为线性对流方程的色散关系 (4.29) 的推广，用 $\omega=g(k)$ 表示这里的色散关系，则群速度 $C(k)$ 由下式定义

$$C(k):=-\mathrm{d}\omega/\mathrm{d}k. \tag{4.86}$$

于是对线性对流方程，上式简化为 $C(k)=a$. 但对于差分格式来说，即便是这种最简单的方程，群速度的表示也不简单. 若记 ω_h 表示 $(\arg\ \lambda)/(\Delta t)$ ，那么由 (4.85) 式，即可给出离散的色散关系

$$\tan\frac{1}{2}\omega_h\Delta t=-\nu\tan\frac{1}{2}k\Delta x.$$

因此，对应于 (4.86)，可推导出差分格式的群速度

$$\begin{aligned}C_h &= \nu\Delta x \sec^2\frac{1}{2}k\Delta x\Big/\left(\Delta t\sec^2\frac{1}{2}\omega_h\Delta t\right)\\ &= a\sec^2\frac{1}{2}k\Delta x\Big/\left(1+\nu^2\tan^2\frac{1}{2}k\Delta x\right)\\ &= \frac{a}{\cos^2\frac{1}{2}\xi+\nu^2\sin^2\frac{1}{2}\xi}.\end{aligned}\tag{4.87}$$

当 $\xi = k\Delta x$ 很小时，为得到离散群速度的误差，可直接把上式展开为 ξ 的幂级数. 然而，利用 ω_h 的展开式 (4.85) 可能更直接更明白，将之对 ξ 微分，立刻可以看到群速度的相对误差是 $\frac{1}{4}(1-\nu^2)\xi^2$. 此误差差不多是相对相速度误差的 3 倍，这显然是因为格式是二阶精度的；还可看出，两种误差之间的关系如此简单，它意味着从群速度我们并不能对格式低频时的行为有更多的理解. 但是，一般来说数值格式的误差会由很多频率组成，因此群速度将是格式行为的一个很好的指南. 特别地，考察频谱中的另一个极端，高频振荡 $\xi=\pi$ 时，可以发现 $C_h = a/\nu^2$，由之也可以看出 $0<\nu<1$ 时，振荡的移动超在主体波前面，另外，与相速度相比，群速度预言的振荡速度将更加精确.

再来考虑稳定性，由于 (4.84)，盒式格式常被认为是无条件稳定的. 但这里必须要求按照正确的方向求解方程组：显然若 $\nu<0$ 时，仍使用递归式 (4.82) 求解，则误差将无限增长；同时，这种计算形式也违背了有关依赖区域上的 CFL 条件. 在实际使用时，对大多数情况来说，当然不难避免这种错误，而格式的无条件稳定性及其简洁的形式使之备受水利工程师的青睐，从而在河流建模中得到广泛的应用.

上面我们已提到了图 4-13(a,b) 中的计算结果，其中仍是求解同样的线性对流方程 (4.33)，初始条件分别是方波和高斯脉冲. 可以看出，由初始间断造成的数值振荡比 Lax-Wendroff 格式时大了许多，然而对光滑初值，二者精度非常相似. 注意到在两种情形下，由于 $|\nu|<1$，所以相误差使得振荡超前于主体波，而且方波产生的振荡向前传播的速率正如群速度所预言的. 这里振荡之所以更严重，是因为该格式没有衰减，而对实际问题，我们可以对 (4.78) 空间差分中用加权平均 $\theta U^{n+1}+(1-\theta)U^n$，其中 $\theta>\frac{1}{2}$，以避免可能出现的振荡.

到目前为止，从上面的叙述中可能并不能明显地看出，为什么此格式与河流模型有密切的联系. 在河流模型中，此格式通常称为 Preissmann 盒式格式 [1]，所谓河流模型，是指通过把横截面面积 $A(x,t)$ 和整体流出量 $Q(x,t)$ 作为独立变量，沿河道积分，推导出的

[1] Preissmann, A.(1961), *Propagation des intumescences dans les canaux et rivières*. Paper presented at the First Congress of the French Association for Computation, Grenoble,France.

河流的 St Venant 方程

$$\begin{aligned} A_t + Q_x &= 0, \\ Q_t + (Q^2/A + \frac{1}{2}gA^2/b)_x &= S, \end{aligned} \tag{4.88}$$

其中 x 表示沿着河流的距离，S 描述了河床斜率和摩擦阻力的作用，g 表示重力常数，b 指河流宽度. 上述第一个方程表示质量守恒，第二个方程是动量方程. 一般来说，河流的这些条件随 x 变化很大. 所以此格式的优点之一就是可以将河流分成很多不同长度的段，在每段中河流的参量差不多是常数. 例如，河流宽度可能变化较大，且与横截面的形状和面积 A 有关，不过这里我们可假设河道截面是长方形的，且其宽度与截面积 A 无关，于是河流高度为 $A/b = h$. 在此假设下，可以简单地计算出通量的雅可比矩阵：利用 h 和流速 $u = Q/A$，可将其表示为

$$A = \begin{pmatrix} 0 & 1 \\ gh - u^2 & 2u \end{pmatrix} \tag{4.89}$$

容易看出它有两个特征值 $u \pm c$，其中 $c = \sqrt{(gh)}$. 对正常的缓缓流动的河流来说，其流速是低于临界速度的，即 $u < c$, 因此其特征速度一正一负，从而方程组需要左右两个边界条件；假定河流从左向右流动，通常给出左边入流处的流入量 Q，以及右端出流处的河流高度 h. 所求的方程组是非线性的，可用牛顿迭代法解之；如前所述，这里所得到的线性方程组是块三对角的，每块是 2×2 矩阵，该方程组可用矩阵形式的 Thomas 算法求解.

然而，尽管该格式非常简洁，且求解非线性方程的过程也相对简单，但我们仍不清楚为什么要用这个隐式格式；并且既然 (4.85) 表明该格式相速度精度并不特别值得称道，为什么它计算洪峰时有很好的精度呢？答案来自洪峰波流的结构，特别是动量方程中那些项的平衡. 动量方程右端表示力的项通常写为 $S = g(S_0 - S_f)$，其中 $S_0 = S_0(x)$ 表示河床斜率，$S_f = S_f(x, A, Q)$ 表示河床摩擦力. 方程中与 g 相乘的三项差不多相互平衡了，由此给出了关系式 $Q = Q_g(x, A)$. 的确，在更简单的模型中，常常把质量守恒方程与一个由试验得出的类似关系式联立，以预测河水的流动情况. 由这个关系式可推知洪峰波的传播速度为 $\partial Q_g/\partial A$，通常此速度会略大于 u，但远比 $u + c$ 小. 于是可依据此速度来选取时间步长（使得 (4.85) 中的 ν 与单位 1 接近，从而给出精确的相速度），同时也充分利用了关于特征速度 $u + c$ 的稳定性条件. 此外，通常在时间平均上选取 $\theta > \frac{1}{2}$，以使伪解波型产生的振荡在 (将在第 5.8 节中讨论) 某种意义下有所衰减.

4.9 蛙跳格式

4-14 所示. 第二个重要的格式称为 **蛙跳格式**(leap-frog scheme), 这是由于格式使用的点如图为在时间方向上使用中心差分, 该格式使用了前面两个时间层上的值; 并且对于中间时间层上的空间差分, 该格式将其"两条腿"伸展开来使用了不相邻的两点上的值. 对于方程 (4.46) 和 (4.47), 蛙跳格式具体形式如下

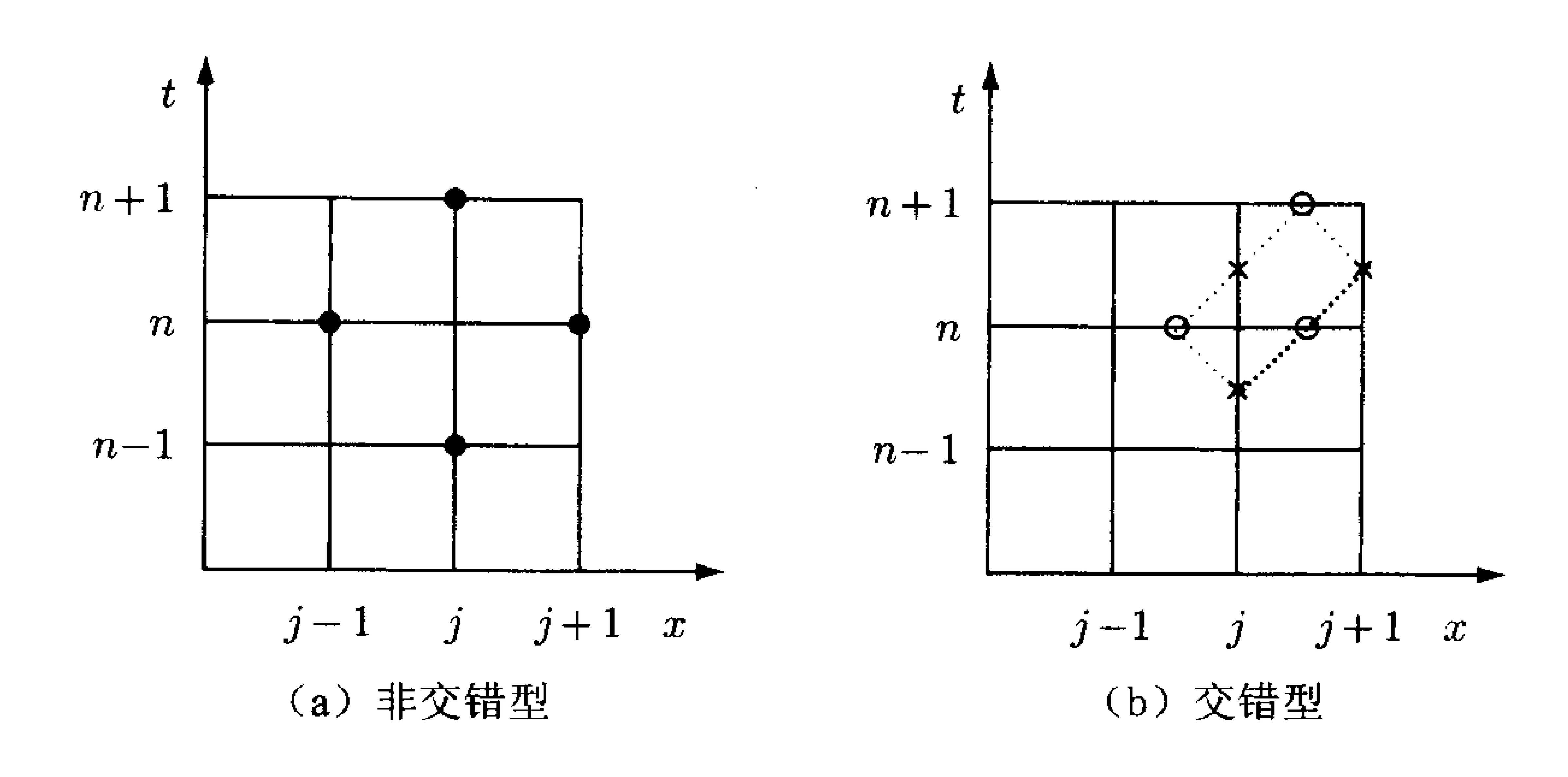


图 4-14 蛙跳格式

$$\frac{U_j^{n+1}-U_j^{n-1}}{2\Delta t}+\frac{f(U_{j+1}^n)-f(U_{j-1}^n)}{2\Delta x}=0, \tag{4.90}$$

或者

$$U_j^{n+1}=U_j^{n-1}-(a\Delta t/\Delta x)\left[U_{j+1}^n-U_{j-1}^n\right]. \tag{4.91}$$

这样, 它是一个显式格式, 但开始时需要特别的技术来启动. 事实上, 初始条件一般只给出了 U^0 的值, 这就需要用特别的方法求出 U^1 的值. 然后, 蛙跳格式才能依次地给出 $U^2, U^3, \ldots$. 这些额外需要的启动值 U^1 可利用任何方便的单步格式得到, 例如可用 Lax-Wendroff 格式求之.

从图 4-14(a) 可以看出, CFL 条件要求 $\nu \leqslant 1$, 这和 Lax-Wendroff 格式要求的一样. 当 $f=au$ 且 a 为常数时, 由通常的傅里叶分析方法, 得到关于 $\lambda(k)$ 的二次方程

$$\lambda^2-1+2\,\mathrm{i}\,\nu\lambda\sin k\Delta x=0, \tag{4.92}$$

其解为

$$\lambda(k)=-\mathrm{i}\,\nu\sin k\Delta x\pm\left[1-\nu^2\sin^2 k\Delta x\right]^{1/2}. \tag{4.93}$$

由于方程的两个解乘积为 -1，所以为了格式稳定，这两个解的模必须都为 1．容易验证，为使这两个根为复数且模对于任意的 k 都相等，必须且仅需 $|\nu| \leqslant 1$，这样，傅里叶分析得到的结果就与 CFL 条件给出的相同；而且满足稳定性条件时格式没有衰减．

蛙跳格式的傅里叶分析导致 $\lambda(k)$ 存在两个值，这是个严重的问题，因为这表明其中必然存在一个伪解波型．产生伪解波型的原因主要是该格式涉及到三个时间层，从而需要额外的初始条件，正是这个额外的初始条件决定了伪解波型的强度． (4.93) 中实部为正的根对应的波型是微分方程真解的一个好的逼近，称之为"真"解波型 λ_T，其辐角为

$$\begin{aligned}\arg\lambda_T &= -\sin^{-1}(\nu\sin k\Delta x)\\ &\sim -\nu\xi\left[1-\frac{1}{6}(1-\nu^2)\xi^2+\cdots\right].\end{aligned}\tag{4.94}$$

注意到这里相误差的主项与 Lax-Wendroff 格式的相同 (参见 (4.40))．另一方面，实部取负的根给出了伪解波型

$$\lambda_S \sim (-1)\left[1+\mathrm{i}\,\nu\xi-\frac{1}{2}\nu^2\xi^2+\cdots\right],\tag{4.95}$$

此波型在时间步间产生振荡并朝错误的方向传播．因此，在实际应用中，必须非常注意不要模拟出这种波型，有时则还需要将之过滤掉．

由 (4.91) 或 (4.92) 可推出蛙跳格式的离散色散关系如下

$$\sin\omega_h\Delta t = -\nu\sin k\Delta x.$$

将上式微分，可得到对应于 (4.86) 的群速度

$$\begin{aligned}C_h &= \nu\Delta x\cos k\Delta x/\Delta t\cos\omega_h\Delta t\\ &= \frac{a\cos k\Delta x}{(1-\nu^2\sin^2 k\Delta x)^{1/2}}.\end{aligned}\tag{4.96}$$

与盒式格式时一样，展开上式或者对 (4.94) 微分，可导出真解波型的群速度误差，也就是说，当 $k\Delta x$ 较小时，群速度误差大约是 (4.94) 给出的相误差的三倍．在另一个极端 $k\Delta x=\pi$ 时，群速度为 $-a$，这即是伪解波型或称寄生波型的群速度．事实上，这等价于在 $k\Delta x=0$ 的极限情况下考虑 (4.93) 中负根的情形，因为当 $k\Delta x$ 在允许范围 $[0,2\pi)$ 内遍历，取值跨过 π 点时，两根的作用将发生改变．

图 4-15 展示了用蛙跳格式求解带有方波和高斯初始值的模型问题时，所得的计算结果，其中我们采用 Lax-Wendroff 格式计算第一步的值．结果清晰地表明振荡是在向左移动．在某些方面，计算结果和盒式格式的有点相似，但这里振荡的移动速度与网格尺寸无关，且不衰减，所以在这种情形下，必须通过某种形式的过滤来抵消之．

我们知道从熟悉的二阶微分方程可以推导出一对一阶方程，当把蛙跳格式应用到这样一对微分方程时，其真正优势才发挥了出来．例如对于波动方程

$$u_{tt} = a^2 u_{xx}, \tag{4.97}$$

其中 a 为常数．若引进变量 $v = u_t$ 以及 $w = -au_x$，则显然它们满足方程组

$$\begin{aligned} v_t + aw_x &= 0, \\ w_t + av_x &= 0. \end{aligned} \tag{4.98}$$

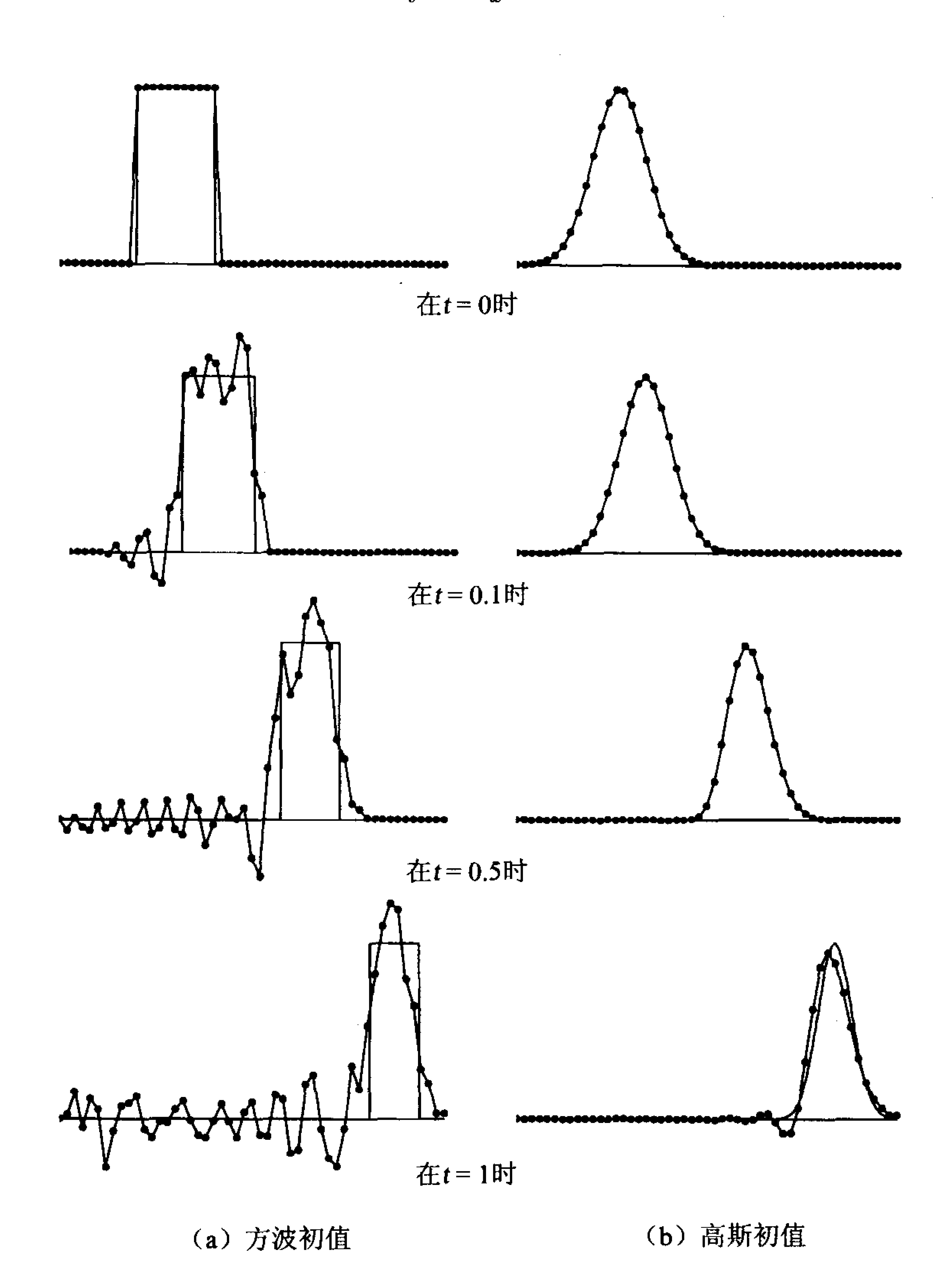


图 4-15　蛙跳格式求解线性对流方程，其中 $\Delta t = \Delta x = 0.02$

考虑到此时各微分项的排列方式，可以使用一种比 (4.91) 更紧凑的交错形式的蛙跳格式，如图 4-14(b) 所示，令 V 和 W 在不同的点取值，所得的交错格式可写为

$$\frac{V_j^{n+1/2}-V_j^{n-1/2}}{\Delta t}+a\frac{W_{j+1/2}^n-W_{j-1/2}^n}{\Delta x}=0, \tag{4.99a}$$

$$\frac{W_{j+1/2}^{n+1}-W_{j+1/2}^n}{\Delta t}+a\frac{V_{j+1}^{n+1/2}-V_j^{n+1/2}}{\Delta x}=0, \tag{4.99b}$$

或

$$\delta_t V+\nu\delta_x W=0,\quad \delta_t W+\nu\delta_x V=0, \tag{4.100}$$

这里使用这种省去了上下标的记号. 由于 a 为常数，可构造傅里叶波型如下

$$(V^{n-1/2},W^n)=\lambda^n \mathrm{e}^{\mathrm{i}kx}(\widehat{V},\widehat{W}), \tag{4.101}$$

其中 $\widehat{V}$ 和 $\widehat{W}$ 为常量. 此波型满足方程 (4.99) 的条件是

$$\begin{pmatrix} \lambda-1 & 2\mathrm{i}\nu\sin\frac{1}{2}k\Delta x \\ 2\mathrm{i}\lambda\nu\sin\frac{1}{2}k\Delta x & \lambda-1 \end{pmatrix}\begin{pmatrix} \widehat{V} \\ \widehat{W} \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.102}$$

这就要求上式中矩阵是奇异的，于是有

$$\lambda^2-2(1-2\nu^2\sin^2\frac{1}{2}k\Delta x)\lambda+1=0, \tag{4.103}$$

其解为

$$\lambda_{\pm}(k)=1-2\nu^2s^2\pm 2\,\mathrm{i}\,\nu s[1-\nu^2s^2]^{1/2}, \tag{4.104}$$

其中 $s=\sin\frac{1}{2}k\Delta x$. 同样地，为使格式稳定， λ_+,λ_- 必须为一对共轭复数，这就要求 $|\nu|\leqslant 1$ ，此时 $|\lambda_{\pm}|=1$. 其相位可由下式给出

$$\begin{aligned}\arg\lambda_{\pm} &= \pm\sin^{-1}\left[2\nu s(1-\nu^2s^2)^{1/2}\right] \\ &\sim \pm\nu\xi\left[1-\frac{1}{24}(1-\nu^2)\xi^2+\cdots\right].\end{aligned} \tag{4.105}$$

注意到将方程 (4.92) 的两根中的 Δx 替换为 $\frac{1}{2}\Delta x$ ，再平方就得到方程 (4.103) 的两根；所以， (4.105) 就相当于 (4.94) ，只是把其中的 ξ 替换成了 $\frac{1}{2}\xi$. 这两种波型都是真解波型，它们以相同的速度分别向左右两个方向移动，正确地描述了波动方程的解的行为. 另外，还可注意到蛙跳格式的精度比盒式格式的高.

把

$$V_j^{n+1/2}=(U_j^{n+1}-U_j^n)/\Delta t,\quad W_{j+1/2}^n=-a(U_{j+1}^n-U_j^n)/\Delta x$$

代入到方程 (4.99) 或 (4.100) 中，得到

$$(\delta_t^2-\nu^2\delta_x^2)U_j^n=0, \tag{4.106}$$

这正是直接用中心差分格式逼近二阶波动方程 (4.97) 中导数而得到的满足相容性条件的最简单的格式. 若从方程中消去 V 或 W，可以发现二者都满足二阶微分方程. 下一节我们将在更广阔的背景下，推导出该格式的一些很诱人的性质.

4.10 哈密顿系统与辛积分格式

交错蛙跳格式对于求解波动方程具有引人注目的效果，这主要是因为以下两个重要的结构上的性质：第一，波动方程 (4.97) 是一个最简单的 *哈密顿*(Hamiltonian)PDE；第二，交错蛙跳格式是一个最常见的 *辛积分格式*(symplectic integration scheme). 过去几年，将两者结合起来逼近常微分方程组的做法已充分展现了其重要性. 但直到最近，这种思想才被引入到 PDE 领域. 下面通过交错蛙跳格式的例子介绍一些主要的相关内容，同时还将说明盒式格式也具有一些同样的性质. 这里主要采用 Leimkuhler 和 Reich(2004) 中的术语和记号.

ODE 的哈密顿系统最早源自哈密顿 1834 年关于动力系统运动方程的公式，后人将之进行了大量推广，并对其主要性质进行了广泛研究. 令 $\mathbf{q}\in\mathbb{R}^d$ 和 $\mathbf{p}\in\mathbb{R}^d$ 分别表示 "位置" 和 "动量" 变量，将它们并在一起记为 $\mathbf{z}$，并记 $\mathcal{H}(\mathbf{q},\mathbf{p})\equiv\mathcal{H}(\mathbf{z}):\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$ 为一光滑哈密顿函数 (Harmiltonian function)，由其可定义 ODE 方程组

$$\dot{\mathbf{z}}\equiv\begin{pmatrix}\dot{\mathbf{q}}\\ \dot{\mathbf{p}}\end{pmatrix}=\mathbf{J}\begin{pmatrix}\mathcal{H}_{\mathbf{q}}\\ \mathcal{H}_{\mathbf{p}}\end{pmatrix}\equiv\mathbf{J}\nabla_{\mathbf{z}}\mathcal{H}, \tag{4.107}$$

其中 *正则结构矩阵* (canonical structure matrix) $\mathbf{J}$ 形式如下

$$\mathbf{J}=\begin{pmatrix}0 & I_d\\ -I_d & 0\end{pmatrix}\ \text{及}\ \mathbf{J}^{-1}=\begin{pmatrix}0 & -I_d\\ I_d & 0\end{pmatrix},$$

这里 I_d 是 d 维单位矩阵. 显然 $\mathcal{H}$ 沿任意一条轨迹都恒为常数，其值表示系统的能量，有时也用 $E(\mathbf{z})$ 表示之. 事实上，考虑任意函数 $\mathcal{G}:\mathbb{R}^{2d}\to\mathbb{R}$, 沿着任意一条轨迹都有

$$\frac{\mathrm{d}\mathcal{G}(\mathbf{z})}{\mathrm{d}t}=(\nabla_{\mathbf{z}}\mathcal{G})^T\dot{\mathbf{z}}=(\nabla_{\mathbf{z}}\mathcal{G})^T\mathbf{J}\nabla_{\mathbf{z}}\mathcal{H}=:\{\mathcal{G},\mathcal{H}\}. \tag{4.108}$$

符号 $\{\mathcal{G},\mathcal{H}\}$ 称为 $\mathcal{G}$ 和 $\mathcal{H}$ 的 *泊松括号*(Poisson bracket). 显然它是反对称的，因而当 $\mathcal{G}=\mathcal{H}$ 时等于零. 另外，只要该表达式为零，那么其对应的 $\mathcal{G}(\mathbf{q},\mathbf{p})$ 的值沿轨迹保持不变，此时，称 $\mathcal{G}$ 为一个 *运动不变量* (constant of the motion). 对于任何哈密顿系统，能量都是运动不变量.

哈密顿系统的最著名的例子是平面单摆，此时 $d=1$ 且 $\mathcal{H}=\frac{1}{2}p^2-(g/L)\cos q$. 由 $\mathcal{H}(q,p)=$ 常数, 可知，其轨迹由方程组 $\dot{q}=p,\dot{p}=-(g/L)\sin q$ 给出. 由此不难导出，在 (q,p)

相平面中这些轨迹是一些常见的闭合曲线，分别以 $p=0,q=2m\pi$ 为中心，彼此被鞍点 $p=0,q=(2m+1)\pi$ 分开.

对于哈密顿系统，比存在不变量更重要的性质是，其轨迹形成的相流的结构性质：例如，在标量情形 $d=1$ 时，它具有*保面积的* (area-preserving) 性质，或者更一般地被称为是*辛的* (symplectic)．为讨论相关内容，还里需要一些定义．一个一般的映射 $\Psi:\mathbb{R}^{2d}\to\mathbb{R}^{2d}$ 被称为辛的，是指对于正则结构矩阵 $\mathbf{J}$，该映射的雅可比矩阵满足

$$\Psi_{\mathbf{z}}^T\mathbf{J}^{-1}\Psi_{\mathbf{z}}=\mathbf{J}^{-1}, \tag{4.109}$$

对于标量情形，容易计算

$$\text{若}\ \Psi_{\mathbf{z}}=\begin{pmatrix} a & b\\ c & d\end{pmatrix}\quad \text{则}\ \Psi_{\mathbf{z}}^T\mathbf{J}^{-1}\Psi_{\mathbf{z}}=\begin{pmatrix} 0 & -ad+bc\\ ad-bc & 0\end{pmatrix},$$

于是，Ψ 是辛的当且仅当 $\det\Psi_z\equiv ad-bc=1$．因此若 Ψ 满足该条件，并且把 $\mathbf{z}\in\Omega\subset\mathbb{R}^2$ 映射到 $\hat{\mathbf{z}}=\Psi(\mathbf{z})\in\hat{\Omega}\subset\mathbb{R}^2$，我们有

$$\int_{\hat{\Omega}}\mathrm{d}\hat{\mathbf{z}}=\int_{\Omega}\det\Psi_{\mathbf{z}}\mathrm{d}\mathbf{z}=\int_{\Omega}\mathrm{d}\mathbf{z},$$

也就是说，映射是保面积的，于是辛性质就是面积保持性质到 $d>1$ 的推广.

为将此概念用到由微分方程积分得到的映射上去，我们用微分几何的语言引入如下定义：函数 $f:\mathbb{R}^{2d}\to\mathbb{R}$ 在方向 $\xi\in\mathbb{R}^{2d}$ 上的*微分 1- 形式* (differential one-form) 定义为

$$\mathrm{d}f(\boldsymbol{\xi}):=\nabla_{\mathbf{z}}f\cdot\boldsymbol{\xi}\equiv\sum_{i=1}^{2d}\frac{\partial f}{\partial z_i}\xi_i. \tag{4.110}$$

而后，对两个函数 f 和 g 可定义*微分 2- 形式*(differential two-form)，或称*楔积*(wedge product), 如下

$$(\mathrm{d}f\wedge\mathrm{d}g)(\boldsymbol{\xi},\boldsymbol{\eta}):=\mathrm{d}g(\boldsymbol{\xi})\mathrm{d}f(\boldsymbol{\eta})-\mathrm{d}f(\boldsymbol{\xi})\mathrm{d}g(\boldsymbol{\eta}). \tag{4.111}$$

特别地，可将 $\mathbf{z}\equiv(\mathbf{q},\mathbf{p})$ 的分量函数 z_i 作为 f 代入到 (4.110) 中，得 $\mathrm{d}z_i(\boldsymbol{\xi})=\xi_i$，再把所有分量写成向量形式 $\mathrm{d}\mathbf{z}\equiv(\mathrm{d}\mathbf{q},\mathrm{d}\mathbf{p})^T=(\mathrm{d}z_1,\mathrm{d}z_2,\ldots,\mathrm{d}z_{2d})^T$. 同样不难看出，若分别把形变变量 $\mathbf{z}=\Psi(\mathbf{z})$ 的分量代入到 (4.110) 中，则得到

$$\mathrm{d}\hat{\mathbf{z}}(\boldsymbol{\xi})=\Psi_{\mathbf{z}}\mathrm{d}\mathbf{z}(\boldsymbol{\xi})\equiv\Psi_{\mathbf{z}}\boldsymbol{\xi}. \tag{4.112}$$

更进一步，把这些分量代入到 (4.111)，可以定义楔积

$$\mathrm{d}\mathbf{q}\wedge\mathrm{d}\mathbf{p}:=\sum_{i=1}^{d}\mathrm{d}q_i\wedge\mathrm{d}p_i. \tag{4.113}$$

事实证明，正是其保持守恒才是哈密顿系统的关键性质.

首先，像标量情形那样，计算可得

$$
\begin{aligned}
\boldsymbol{\xi}^T J^{-1}\boldsymbol{\eta} &= (\mathrm{d}\mathbf{q}^T(\boldsymbol{\xi}), \mathrm{d}\mathbf{p}^T(\boldsymbol{\xi}))J^{-1}(\mathrm{d}\mathbf{q}(\boldsymbol{\eta}), \mathrm{d}\mathbf{p}(\boldsymbol{\eta}))^T \\
&= \sum_{i=1}^{d}[\mathrm{d}p_i(\boldsymbol{\xi})\,\mathrm{d}q_i(\boldsymbol{\eta}) - \mathrm{d}q_i(\boldsymbol{\xi})\,\mathrm{d}p_i(\boldsymbol{\eta})] \\
&= \sum_{i=1}^{d}\mathrm{d}q_i \wedge \mathrm{d}p_i \equiv \mathrm{d}\mathbf{q}\wedge\mathrm{d}\mathbf{p}. \qquad (4.114)
\end{aligned}
$$

然后，对 (4.109) 式左乘 $\boldsymbol{\xi}^T$ 右乘 $\boldsymbol{\eta}$，将结果与 (4.112) 和 (4.114) 联立所得的结果对照，可立即推出以下结论： $(\mathbf{q}, \mathbf{p})$ 到 $(\hat{\mathbf{q}}, \hat{\mathbf{p}})$ 的映射是辛的当且仅当

$$
\mathrm{d}\hat{\mathbf{q}}\wedge\mathrm{d}\hat{\mathbf{p}} = \mathrm{d}\mathbf{q}\wedge\mathrm{d}\mathbf{p}. \qquad (4.115)
$$

哈密顿系统的相流是辛的，这个基本结果可从 (4.109) 直接推出，而 (4.115) 对描述相流是至关重要的.

用于逼近 ODE 方程组的数值方法，若保持这些性质，则称之为 *辛积分格式*(symplectic integration scheme)，或者更一般地， *几何积分子*(geometric integrator)(参见 Hairer, Lubich 和 Wanner(2002))．辛积分格式中最简单的当属与蛙跳格式一样具有交错结构的格式了，为简单起见，我们先来讨论标量，即 $d=1$ 的情形，此时交替使用下面两个方程

$$
\begin{aligned}
q^{n+1} &= q^n + \Delta t\mathcal{H}_p(q^n, p^{n+1/2}) \\
p^{n+1/2} &= p^{n-1/2} - \Delta t\mathcal{H}_q(q^n, p^{n+1/2}). \qquad (4.116)
\end{aligned}
$$

如果与单摆的情形一样， $\mathcal{H}_p$ 只依赖于 p， $\mathcal{H}_q$ 只依赖于 q，则上式为显式格式；而对更一般的情形，其为隐式格式．在任何一种情况下，对上式求微分，得到

$$
\begin{aligned}
\mathrm{d}q^{n+1} &= \mathrm{d}q^n + \Delta t\left[\mathcal{H}_{pq}\,\mathrm{d}q^n + \mathcal{H}_{pp}\,\mathrm{d}p^{n+1/2}\right] \\
\mathrm{d}p^{n+1/2} &= \mathrm{d}p^{n-1/2} - \Delta t\left[\mathcal{H}_{qq}\,\mathrm{d}q^n + \mathcal{H}_{qp}\,\mathrm{d}p^{n+1/2}\right], \qquad (4.117)
\end{aligned}
$$

这里我们省略了 (4.116) 式中出现的哈密顿函数的变量．现把第一个方程两边与 $\mathrm{d}p^{n+1/2}$ 做楔积，并只把含有 $\mathrm{d}q^n$ 的项中的 $\mathrm{d}p^{n+1/2}$ 用第二个方程代入；由反对称性可知 $\mathrm{d}q^n\wedge\mathcal{H}_{qq}\,\mathrm{d}q^n$ 和 $\mathcal{H}_{pp}\,\mathrm{d}p^{n+1/2}\wedge\mathrm{d}p^{n+1/2}$ 为零；因此，消去零项得到

$$
\begin{aligned}
\mathrm{d}q^{n+1}\wedge\mathrm{d}p^{n+1/2} &= \mathrm{d}q^n\wedge\left[\mathrm{d}p^{n-1/2} - \Delta t\mathcal{H}_{qp}\,\mathrm{d}p^{n+1/2}\right] \\
&\quad +\Delta t\mathcal{H}_{pq}\,\mathrm{d}q^n\wedge\mathrm{d}p^{n+1/2}. \qquad (4.118)
\end{aligned}
$$

含 Δt 的两个项相互抵消，于是得到离散形式的辛性质

$$
\mathrm{d}q^{n+1}\wedge\mathrm{d}p^{n+1/2} = \mathrm{d}q^n\wedge\mathrm{d}p^{n-1/2}. \qquad (4.119)
$$

若对维数 $d>1$ 的情形重复上述整个过程，也可得到同样的结果．这是因为，从 (4.111)

和 (4.113) 的定义中容易看出，对任意的矩阵 A 有

$$\mathbf{da} \wedge (A\,\mathbf{db}) = (A^T\mathbf{da}) \wedge \mathbf{db},$$

于是，若 A 对称且 $\mathbf{a} = \mathbf{b}$，则由楔积的反对称性可知其结果为零，这样就可同样地消去一些项了.

在有关 ODE 的文献中，这种交错蛙跳格式通常被称为 Störmer-Verlet 方法，它与常用的不对称欧拉方法的区别仅在于上标的标注. 在解决哈密顿 ODE 方程组的长时间积分问题时，该方法具有很好的有效性，这在已引用的文献中都可看到.

把这些思想应用到 PDE 上是最近的事情，其中有几种不同的方法：一种方法是先对空间离散得到一组哈密顿 ODE，再对之直接用上述方法. 为实现第一步，人们日益感兴趣的是用无网格方法或者质点法 (particle method)，但本书没有涉及也不准备涉及质点法；另外，也可先在空间上离散，然后用“线法” (method of line) 在时间上积分，但这里也不考虑之. 一个更基本的公式是由 Bridges[1] 建立的，该公式导出一个*多辛* (multi-symplectic) PDE，将 (4.107) 推广为如下形式

$$\mathbf{K}\mathbf{z}_t + \mathbf{L}\mathbf{z}_x = \nabla_{\mathbf{z}} S(\mathbf{z}), \tag{4.120}$$

其中 $\mathbf{K}$ 和 $\mathbf{L}$ 为反对称常数矩阵. 然而，这两个矩阵及其线性组合通常是奇异的，而且如何将所给系统写成这种形式并不显然. 因此，对波动方程，我们将用一个更直接的方法推广 (4.97) 和 (4.98).

假定存在哈密顿函数 $\mathcal{H}(u,v)$, 它是一个关于空间变量的积分，被积函数是关于 u 和 v 及其导数的函数. 然后，为导出哈密顿 PDE，我们定义 $\mathcal{H}$ 的*变分导数* (variational derivative). 例如，考虑

$$\mathcal{H}(u,v) = \int E(x,t)\mathrm{d}x \equiv \int \left[f(u) + g(u_x) + \frac{1}{2}v^2 \right] \mathrm{d}x, \tag{4.121}$$

这里并未指定方程成立的以及 u 和 v 定义的具体区间；被积函数 $E(x,t)$ 称为能量密度. 一个泛函 $\mathcal{G}(u)$ 的变分导数由以下关系式定义

$$\int \delta_u \mathcal{G}(u)(\delta u)\mathrm{d}x = \lim_{\epsilon \to 0} \frac{\mathcal{G}(u+\epsilon\delta u) - \mathcal{G}(u)}{\epsilon}.$$

利用上式对 (4.121) 求变分导数，假定边界条件保证了任何边界项都为零，就有

$$\begin{aligned} \int \delta_u \mathcal{H}(u,v)(\delta u)\mathrm{d}x &= \lim_{\epsilon \to 0} \epsilon^{-1} \int [f(u+\epsilon\delta u) - f(u) + g((u+\epsilon\delta u)_x) - g(u_x)]\,\mathrm{d}x \\ &= \int [f'(u)\delta u + g'(u_x)(\delta u)_x]\,\mathrm{d}x = \int [f'(u) - \partial_x g'(u_x)]\delta u\,\mathrm{d}x. \end{aligned}$$

---

[1] Bridges, T.J.(1997), Multi-symplectic structure and wave propagation, *Math. Proc. Camb. Philos. Soc.* **121**, 147-90.

比较上式两边，推知

$$\delta_u\mathcal{H}(u,v)=f'(u)-\partial_x g'(u_x). \tag{4.122}$$

由此得到的哈密顿 PDE 为

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix}=\begin{pmatrix} 0 & +1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \delta_u\mathcal{H} \\ \delta_v\mathcal{H} \end{pmatrix}. \tag{4.123}$$

亦即

$$u_t=v, \quad v_t=\partial_x g'(u_x)-f'(u). \tag{4.124}$$

另外，从这些方程中可以推出形如 $E_t+F_x=0$ 的一种局部能量守恒律：对 (4.121) 式的能量密度求导，并将 (4.124) 代入，消去并且合并同类项，得

$$\begin{aligned} E_t &= f'(u)v+g'(u_x)v_x+v\left[\partial_x g'(u_x)-f'(u)\right] \\ &= \left[vg'(u_x)\right]_x =: -F_x. \end{aligned} \tag{4.125}$$

变量 $F(x,t)=-vg'(u_x)$ 称为能量通量.

例如，令 $f=0$ 且 $g(u_x)=\frac{1}{2}(au_x)^2$，其中 a 为常数. 那么 (4.124) 化为

$$u_t=v, \quad v_t=a^2u_{xx}, \tag{4.126}$$

这恰等价于二阶波动方程 (4.97)．若记 $w=-au_x$，则可得到一对一阶方程 (4.98)，我们在第 4.9 节曾用交错差分格式求解之. 而且，因为 $vg'(u_x)=va^2u_x=-avw$，局部能量守恒律变为

$$\left[\frac{1}{2}v^2+\frac{1}{2}w^2\right]_t+[avw]_x=0, \tag{4.127}$$

上式也可由 (4.98) 直接推导出. 这里要证明的是，正是这种局部性质被交错蛙跳格式以离散形式在整个计算过程中得到保持. 而这也可被认为是该格式辛性质的最简单的推论，对应了 ODE 情形能量在运动中保持不变的性质. 为得到楔积关系式 (4.119)，则需要对空间变量积分或求和，已超出了本书的范围.

利用 (4.100) 中引进的简约符号以及 (4.79) 式中定义的平均算子，可得

$$(\mu_t V)(\delta_t V)\equiv\frac{1}{2}(V^+ + V^-)(V^+ - V^-)=\delta_t(\frac{1}{2}V^2).$$

于是，由 (4.100) 中的两个方程可得到

$$\delta_t[\frac{1}{2}(V^2+W^2)]+\nu[(\mu_t V)(\delta_x W)+(\mu_t W)(\delta_x V)]=0. \tag{4.128}$$

上式中第一项显然表示能量在某一时间步上的差分，事实上还除了 Δt，这在第二项的因子 ν 中体现；式中第二项与 a， V， W 和 $(\Delta x)^{-1}$ 都成正比，但是，由此并不能明显地看出该项就是能量通量的差分. 为表明其确实就是能量的差分，需要恢复其上下标，返

过来考虑格式的完整形式 (4.99)，并参考图 4-14(b) 的交错型网格．先写出 (4.128) 式中的能量差分 ΔE, 其具体形式如下

$$\Delta E=\frac{1}{2}\left[(V_j^{n+1/2})^2+(W_{j+1/2}^{n+1})^2\right]-\frac{1}{2}\left[(V_j^{n-1/2})^2+(W_{j+1/2}^{n})^2\right]. \tag{4.129}$$

然后，也可写出相应的通量差分 ΔF，消去项 $V_j^{n+1/2}W_{j+1/2}^n$ 并整理，得

$$\begin{aligned}a^{-1}\Delta F &\equiv \frac{1}{2}\left(V_j^{n+1/2}+V_j^{n-1/2}\right)\left(W_{j+1/2}^n-W_{j-1/2}^n\right)\\ &\quad+\frac{1}{2}\left(W_{j+1/2}^{n+1}+W_{j+1/2}^n\right)\left(V_{j+1}^{n+1/2}-V_j^{n+1/2}\right)\\ &= \frac{1}{2}\left[V_j^{n-1/2}W_{j+1/2}^n+V_{j+1}^{n+1/2}W_{j+1/2}^n+V_{j+1}^{n+1/2}W_{j+1/2}^{n+1}\right]\\ &\quad-\frac{1}{2}\left[V_j^{n+1/2}W_{j+1/2}^{n+1}+V_j^{n+1/2}W_{j-1/2}^n+V_j^{n-1/2}W_{j-1/2}^n\right].\end{aligned} \tag{4.130}$$

上式看上去极其复杂，但从图 4-14(b) 中却可得到简单明了的解释：若连结每个乘积中 V 取值的点和 W 取值的点，可以发现第一个括号中给出了从 $(j,n-\frac{1}{2})$ 绕到 $(j+\frac{1}{2},n+1)$ 经过的 '斜置' 矩形的两条边，另一个括号中则给出了该矩形的另两条边（此斜置矩形如图中点线所示）.

由差分表达式 (4.129) 和 (4.130) 可看出，(4.128) 确实就是所需要的局部守恒形式，然而最好将之写成以下形式

$$\Delta E\Delta x+\Delta F\Delta t=0; \tag{4.131}$$

因为此式正表示守恒律 $E_t+F_x=0$ 在上述斜置矩形上的积分，严格地说，是利用高斯散度定理得到的 $F\mathrm{d}t-E\mathrm{d}x$ 沿周边的线积分．例如，将被积函数 $\frac{1}{2}v^2$ 从 $(j+\frac{1}{2},n+1)$ 至 $(j-\frac{1}{2},n)$，再到 $(j+\frac{1}{2},n)$ 积分给出了 (4.129) 中 $(V)^2$ 项的差分；而再继续沿边界积分，直到返回 $(j+\frac{1}{2},n+1)$，这一过程对整体没有贡献，因为此时 x 的静位移为零．相似地，将被积函数 avw 从 $(j+\frac{1}{2},n)$ 到 $(j+\frac{1}{2},n+1)$ 积分，给出 (4.130) 中的项 $\frac{1}{2}V_{j+1}^{n+1/2}(W_{j+1/2}^n+W_{j+1/2}^{n+1})$，而在前面的几段边上，从 $(j-\frac{1}{2},n)$ 到 $(j+\frac{1}{2},n)$ 积分，得到式中第一项和最后一项的和为 $\frac{1}{2}V_j^{n-1/2}(W_{j+1/2}^2-W_{j-1/2}^n)$．

已有论文显示盒式格式也具有多辛性质 [1]．作为本节的结束，这里只指出被用于求解波动方程组时，盒式格式满足一种简单形式的能量守恒律．利用 (4.80) 的简约符号，可将该格式写为

$$\delta_t\mu_x V+\nu\delta_x\mu_t W=0,\quad \delta_t\mu_x W+\nu\delta_x\mu_t V=0. \tag{4.132}$$

---

[1] Zhao, P.F. and Quin, M.Z. (2000), Multisymplectic geometry and multisymplectic Preissmann scheme for the KdV equation, *J. Phys. A* **33**,3613-26.

然后，使用与推导 (4.128) 同样的方法，可以推出

$$\delta_t \frac{1}{2}\left[(\mu_x V)^2+(\mu_x W)^2\right]+\nu\left[(\mu_t\mu_x V)(\delta_x\mu_t W)+(\mu_t\mu_x W)(\delta_x\mu_t V)\right]=0. \tag{4.133}$$

容易验证

$$\begin{aligned}(\mu_x A)(\delta_x B)+(\mu_x B)(\delta_x A) &\equiv \frac{1}{2}\left[(A^+ + A^-)(B^+ - B^-)\right.\\ &\quad \left.+(B^+ + B^-)(A^+ - A^-)\right]\\ &= A^+B^+ - A^-B^- \equiv \delta_x(AB).\end{aligned} \tag{4.134}$$

所以可以推出

$$\delta_t \frac{1}{2}\left[(\mu_x V)^2+(\mu_x W)^2\right]+\nu\delta_x\left[(\mu_t V)(\mu_t W)\right]=0, \tag{4.135}$$

这是原始网格盒子区域上自然的能量守恒律.

4.11 相误差和振幅误差的比较

现在我们再次使用傅里叶分析，并由此比较本章引进的各种差分格式. 如前所知，当 $\omega=-ak$ 时，傅里叶波型 $u(x,t)=\mathrm{e}^{\mathrm{i}(kx+\omega t)}$ 是常系数微分方程 $u_t+au_x=0$ 的精确解. 此波型的振幅不会衰减，而在每一时间步中其相位的增长量为 $\omega\Delta t=-ak\Delta t$. 此方程的数值格式具有形如 $\lambda^n e^{\mathrm{i}kj\Delta x}$ 的解，其中 $\lambda(k)$ 是 k,Δt 和 Δx 的函数. 在每一时间步，该波型变为原来的 λ 倍，其中 λ 为复数. λ 的模决定了格式的稳定性：若 $|\lambda|>1+O(\Delta t)$ ，格式不稳定，若 $|\lambda|<1$ ，则对应的波型将有衰减. 数值解对精确解的相对相位是指下面的比例

$$\frac{\arg\lambda}{-ak\Delta t}=-\frac{\arg\lambda}{\nu\xi},$$

其中 $\xi=k\Delta x$ 并且 $\nu=a\Delta t/\Delta x$.

图 4-16 显示了 $|\lambda|$ 作为 ξ 的函数的图像，对应于我们讨论过的四种格式的其中两种，图 4-17 则 给出了这四种格式的相对相位的图像. 当 ξ 很小时，这些量都很接近于单位 1，而这些量对于 1 的偏离程度则是格式中数值误差的一种量度. 这里我们给出的是 $0\leqslant\xi\leqslant\pi$ 区间上的图像，尽管只有区间 $[0,\frac{1}{2}\pi]$ 上的相位误差才有比较重要的意义，其原因是更高频率的振荡一般不能在网格上较好地表示出来. 当 $\nu<1$ 时，盒式格式和蛙跳格式没有衰减，所以它们的 $|\lambda|$ 的图像被省略. 对迎风格式和 Lax-Wendroff 格式，分别给出了一条 $\nu>1$ 时的变化曲线，以展示不稳定时的情形.

图 4-17 清楚地显示了 $\nu<1$ 时，盒式格式相位有所超前， Lax-Wendroff 格式和蛙跳格式的相位有所滞后. 当 ξ 值较小时， Lax-Wendroff 格式和蛙跳格式的相对相位几乎一

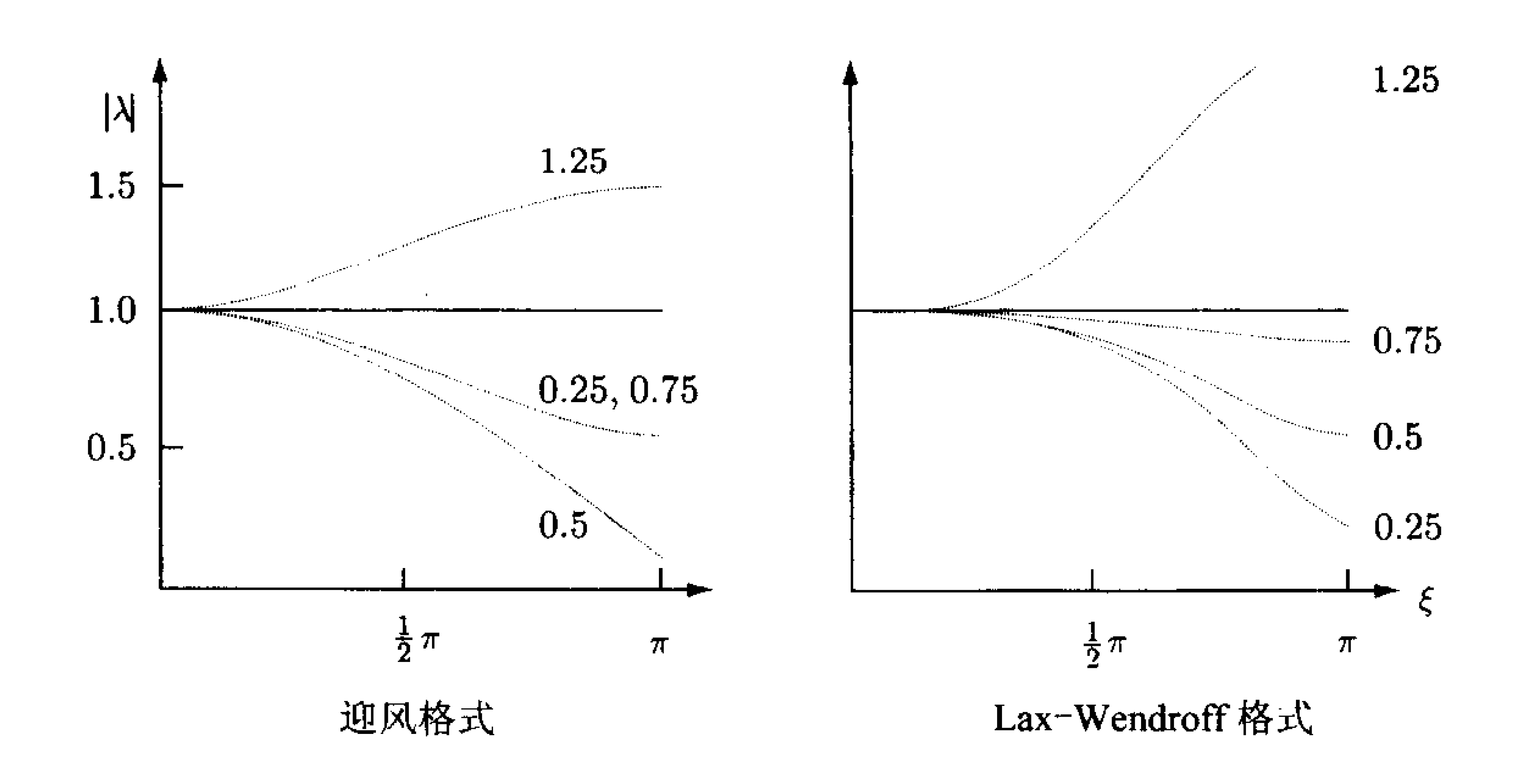


图 4-16 傅里叶波型的增长因子，自变量为 ξ，其中 $\nu = 0.25, 0.5, 0.75$ 和 1.25．注意到迎风格式的相应于 $\nu = 0.25$ 和 $\nu = 0.75$ 的曲线重合

样，而盒式格式的相对相位误差之模几乎是它们的一半，但除了 ν 接近于单位 1 的情形，盒式格式的相对相位的增长速度比前两者大得多．同时也可注意到，蛙跳格式的相对相位受 ν 的影响很小．在图 4-18 中，给出了迎风格式和 Lax-Wendroff 格式的复增长因子 λ 作为 ξ 的函数在复平面内的极坐标轨迹．真解对应的因子为 $e^{-i\nu\xi}$，其在复平面的轨迹是单位圆．图中给出了 $0 \leqslant \xi \leqslant \pi$ 时的图像，并且每隔 $\frac{1}{8}\pi$ 都标记了一个点．所给的曲线是 $\nu = 0.75$ 时情形．

对于盒式格式和蛙跳格式，这样的轨迹图像用处不大，因为此时所有的点都仍在单位圆上；不过，我们可画出其相对群速度 C_h/a，两者各自对应的公式是 (4.87) 及 (4.96)．图 4-19 给出了不同 ν 值时，这两种格式的群速度的情况．

4.12　边界条件与守恒性质

除了与微分方程对应的边界条件外，Lax-Wendroff 格式和非交错蛙跳格式都需要额外的边界条件．事实上，为计算 U_j^{n+1} 的值，需要上一时间层上的 $j-1$ 和 $j+1$ 两点上的值，而这只有当 x_j 为内点时，这些所需的值才都已经存在．因此，对于定义在区间 $(0,1)$ 上的对流方程，其中 $a > 0$，格式不仅需要微分方程提供的左边界条件，而且还需要一个右边界条件．对于方程组，需要对每个方程的差分格式都给出两个边界条件，而不仅仅是通常所给的微分方程组的一个边界条件．盒式格式与交错型蛙跳格式的一个优点就是它们不需要额外的边界条件．事实上，对于波动方程，若左右各给一个边界条件，分别给出 v 和 w 的某一值，则交错蛙跳格式就可以顺利地进行下去．

于是，下面就可以提出这样的问题：应如何推导出这些所需的额外的边界条件呢？所给出的这些条件对格式的性质将会产生怎样的影响呢？一般地说，可用两种方法给出这些额外的边界条件，也有两种方法用以检测所给边界条件对格式的稳定性和精度的影响. 差分格式和微分方程组的边界条件之所以有差别，主要是因为后者一部分特征线指向了边界外部. 因此，最好通过逼近相应的特征线方程来推导这些额外的数值边界条件. 例如，对于对流方程，就可以利用迎风格式 (4.20) 来计算 Lax-Wendroff 格式和非交错蛙跳格式的最后一个点上的值.

但是对于方程组的情形，这种办法需要求出方程组的正规特征分解 (characteristic normal form)，这就需要求出边界点上雅可比矩阵的全部特征值和特征向量. 一个更简易的替代方法是将每个未知量的一阶差分 (或更高阶差分) 简单地置为零，这样，右边界处的一个典型的选择是令 $U_J^n = U_{J-1}^n$.

分析这些额外边界条件影响的最精确的方法，是考察这些条件对传向边界的波的反

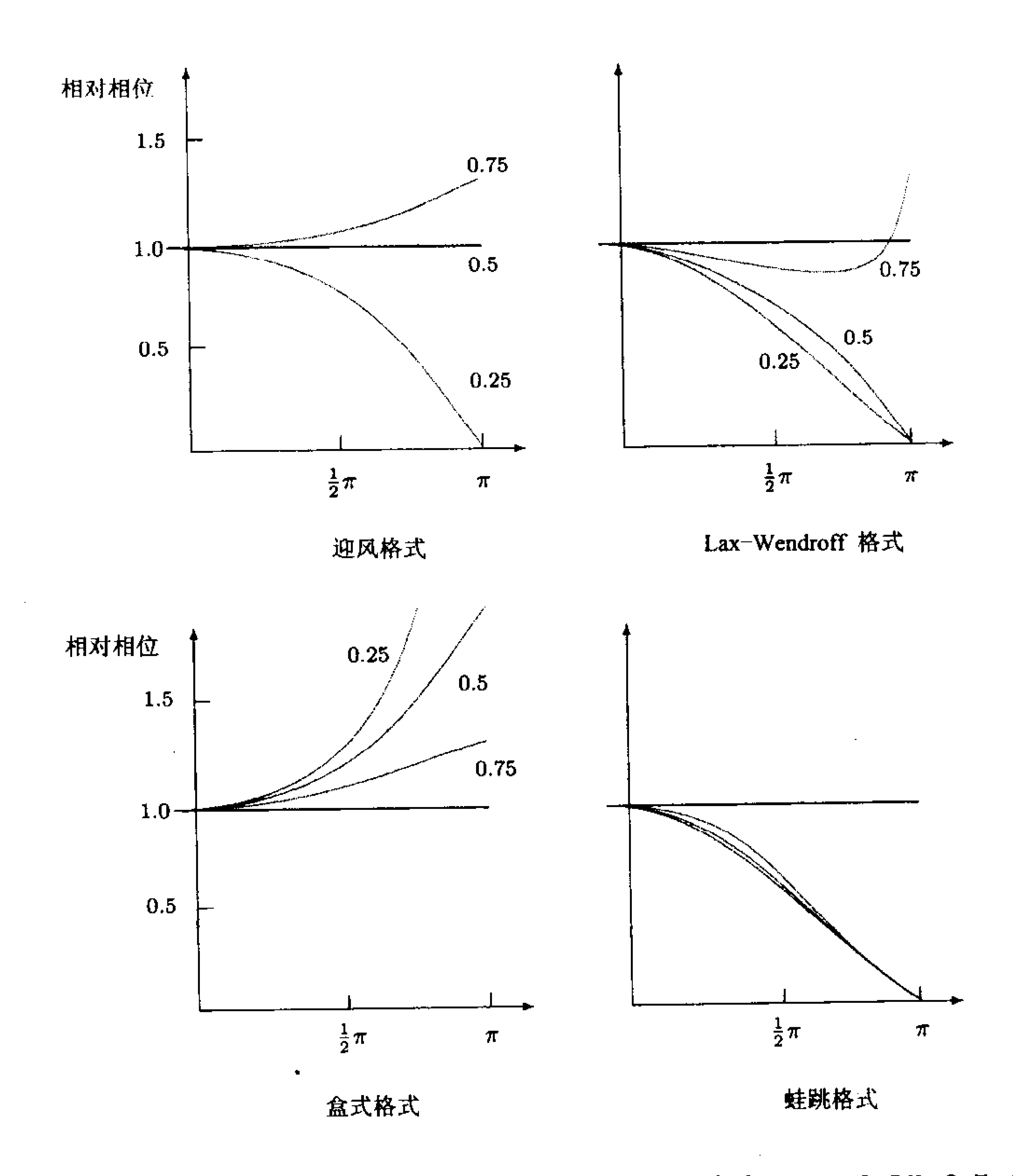


图 4-17 傅里叶波型的相对相位，自变量为 ξ，其中 $\nu = 0.25, 0.5$ 和 0.75; 蛙跳格式的三条曲线非常接近，它们随 ν 增大而增大

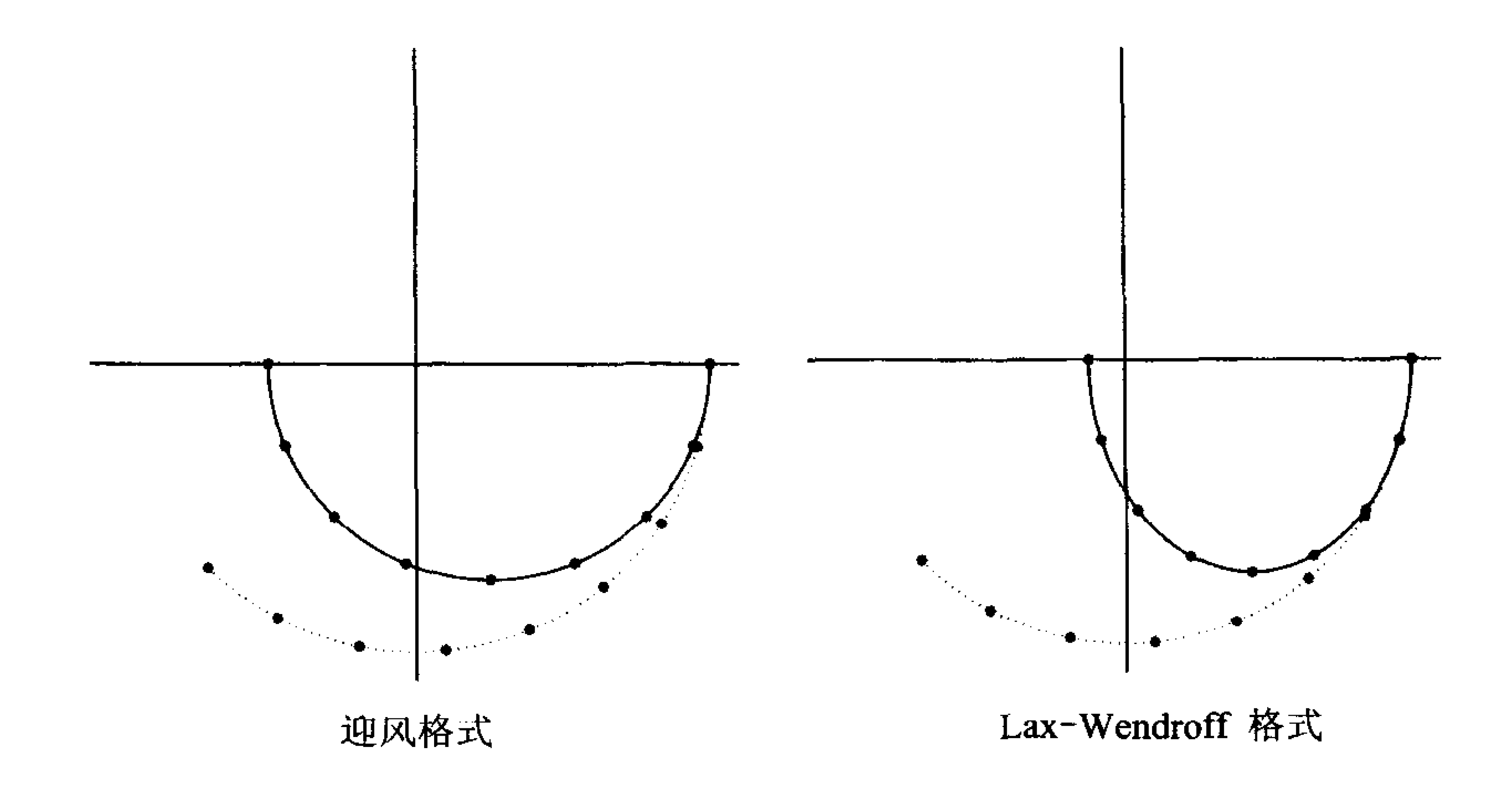


图 4-18　$\nu = 0.75$ 时复增长因子的轨迹；点线对应于真解

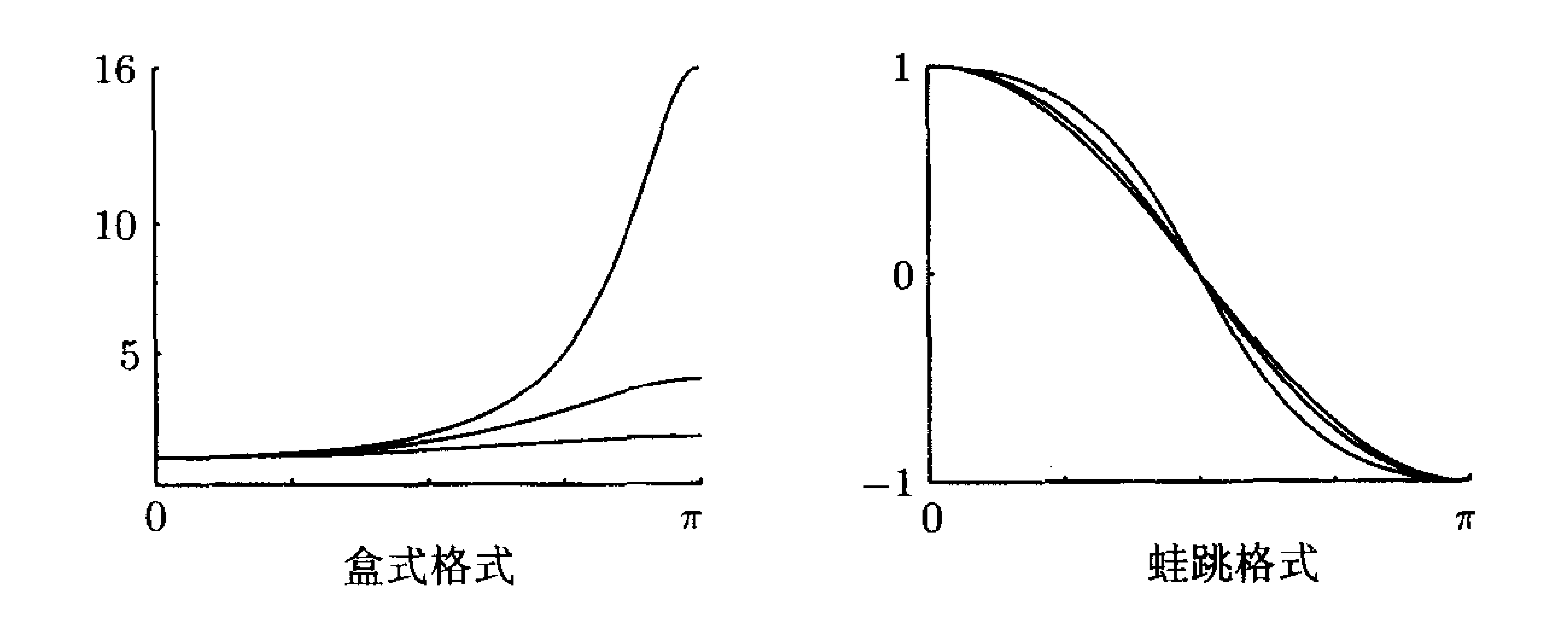


图 4-19　盒式格式和蛙跳格式的群速度，其中 $\nu = 0.25, 0.5, 0.75$

射与传播的影响作用. 我们已经知道，非交错蛙跳格式存在伪解波型，因此，应当保证真解波型传播到边界时，不能产生反射到区域内部中的伪解波型. 这样的分析非常有效，但也相当复杂，完整的讨论超出了本书的范围. 本节后一部分将用这种方法分析一个特定的例子.

另一个方法是选择边界条件时应尽量使差分格式具有与微分方程相似的守恒性质. 事实上，若微分方程 (4.46) 定义在单位区间内，且满足 $u(0,t) = 0$ ，则可得到

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 u(x,t)\mathrm{d}x = \int_0^1 u_t\mathrm{d}x = -\int_0^1 f_x\mathrm{d}x = f(u(0,t)) - f(u(1,t)). \tag{4.136}$$

对于两步 Lax-Wendroff 格式 (4.65)，对应的关系应是

$$\Delta x\sum_{1}^{J-1}(U_j^{n+1} - U_j^n) = \Delta t\left[f(U_{1/2}^{n+1/2}) - f(U_{J-1/2}^{n+1/2})\right]. \tag{4.137}$$

选择两边的边界条件时应该使 (4.137) 是 (4.136) 的一个很好的近似. 下章将对这种思想做更多的阐述.

这里给出一个有关波在边界上传播和反射的例子, 对于方程组 $u_t + av_x = 0, v_t + au_x = 0$, 其中 a 是常数, 给定左边 $x = 0$ 的边界条件, 并用非交错蛙跳格式逼近它. 上述波动方程组有两组解波型, 一组对应于左行波, 记为

$$u_R = \mathrm{e}^{-\mathrm{i}\,k(x-at)}, \quad v_R = \mathrm{e}^{-\mathrm{i}\,k(x-at)}; \tag{4.138}$$

另一组对应于右行波, 记为

$$u_L = \mathrm{e}^{\mathrm{i}\,k(x+at)}, \quad v_L = -\mathrm{e}^{\mathrm{i}\,k(x+at)}. \tag{4.139}$$

若 $x = 0$ 处的边界条件是 $v = 0$, 那么形如 $u = A(u_R + u_L), v = A(v_R + v_L)$ 的表达式都满足差分方程和边界条件. 现在我们对蛙跳式差分方程给出适当的边界条件, 构造出具有相似形式的解.

对差分方程, 我们不能假定其波型解具有指数形式, 对某对 (λ, μ), 而应假定它们具有以下形式

$$U_j^n = \lambda^n \mu^j \widehat{U}, \quad V_j^n = \lambda^n \mu^j \widehat{V}, \tag{4.140}$$

将之代到如下差分格式中

$$U_j^{n+1} - U_j^{n-1} + \nu(V_{j+1}^n - V_{j-1}^n) = 0, \tag{4.141a}$$

$$V_j^{n+1} - V_j^{n-1} + \nu(U_{j+1}^n - U_{j-1}^n) = 0, \tag{4.141b}$$

其中 $\nu = a\Delta t/\Delta x$. 从而得到代数方程组

$$\begin{pmatrix} (\lambda^2 - 1)\mu & \nu\lambda(\mu^2 - 1) \\ \nu\lambda(\mu^2 - 1) & (\lambda^2 - 1)\mu \end{pmatrix} \begin{pmatrix} \widehat{U} \\ \widehat{V} \end{pmatrix} = \mathbf{0}. \tag{4.142}$$

为使上式具有非平凡解, (λ, μ) 必须使其系数矩阵行列式为零, 即满足以下方程

$$(\lambda^2 - 1)^2 \mu^2 - \nu^2 \lambda^2 (\mu^2 - 1)^2 = 0. \tag{4.143}$$

若固定 μ , (4.142) 和 (4.143) 是傅里叶分析所得结果的推广. 例如, 对于求解波动方程的交错蛙跳格式, 相应的方程为 (4.102) 和 (4.103) . 然而, 为了研究边界上的情形, 更好的方式是固定 λ (从而固定了关于 t 的变差) 而考虑由 μ 值确定的空间波型. 可以看出 (4.143) 有四个解, 它们排列得非常对称. 事实上, (4.143) 可改写为

$$\mu^4 - 2\left[1 + \frac{(\lambda^2 - 1)^2}{2\nu^2\lambda^2}\right]\mu^2 + 1 = 0, \tag{4.144}$$

把 μ^2 看成未知量, 上式是一个二次方程, 其根乘积是 1 ; 若记 $\gamma = (\lambda^2 - 1)/\nu\lambda$, 则可得

$$\mu = \pm\left[1 + \frac{1}{2}\gamma^2 - \gamma(1 + \frac{1}{4}\gamma^2)^{1/2}\right]^{\pm 1/2}. \tag{4.145}$$

现考虑能在网格上得到很好逼近的低频波型，与 (4.138) 比较，令 $\lambda \approx \mathrm{e}^{\mathrm{i}ka\Delta t} \approx 1 + ika\Delta t$，于是有 $\gamma \approx 2ik\Delta x.\gamma$ 较小时，由 (4.145) 得到

$$\mu \sim \pm(1-\gamma)^{\pm 1/2}, \quad 当\ \gamma \to 0. \tag{4.146}$$

因此 $(1-\gamma)^{1/2} \approx \mathrm{e}^{-\mathrm{i}k\Delta x}$，并且它对应于一个右行真解波型，$(1-\gamma)^{-1/2}$ 对应于一个左行真解波型，而 $-(1-\gamma)^{\pm 1/2}$ 则对应于蛙跳格式的两个伪解波型. 对每个 λ，记这些相应的根为 $\mu_{RT}, \mu_{LT}, \mu_{RS}, \mu_{LS}$，它们之间的关系为

$$\mu_{RS} = -\mu_{RT}, \quad \mu_{LS} = -\mu_{LT} = -1/\mu_{RT}. \tag{4.147}$$

只要再给定时间变差，我们就可以由它们的组合构造解.

从 (4.142) 式可知，对于每个 μ，相应的特征向量满足

$$\widehat{U} : \widehat{V} = (1-\mu^2) : \mu\gamma, \tag{4.148}$$

并且若已给定两左行波的振幅，则由两个所需的边界条件就可以决定右行波的振幅. 假定左行真解波型的振幅为 1，并且不存在任何伪解波型的噪声；那么对方程 (4.141)，我们有如下形式的解

$$\begin{pmatrix} U_j^n \\ V_j^n \end{pmatrix} = \lambda^n \left\{ \mu_{LT}^j \begin{pmatrix} \widehat{U}_{LT} \\ \widehat{V}_{LT} \end{pmatrix} + \alpha\mu_{RT}^j \begin{pmatrix} \widehat{U}_{RT} \\ \widehat{V}_{RT} \end{pmatrix} + \beta\mu_{RS}^j \begin{pmatrix} \widehat{U}_{RS} \\ \widehat{V}_{RS} \end{pmatrix} \right\}, \tag{4.149}$$

其中 α 和 β 要由边界条件确定. 与真解最匹配的结果是令上式中的 $\alpha = 1, \beta = 0$，此时不存在由于波的反射而产生的伪解波型； Mastsuno[1] 构造的关于数值天气预报的格式就取到了这样的效果. 下面的讨论与处理格式 (2.109) 时的讨论相似，我们曾用这种方法处理热流的诺伊曼边界条件，在第 2.14 节还用同样的方法推导了一些守恒性质. 令

$$U_0^n - U_1^n = 0, \quad V_0^n + V_1^n = 0. \tag{4.150}$$

上式第二个条件是初始边界条件 $v = 0$ 的近似，而第一个则是将差分置为零. 把 (4.149) 正则化，使 $\widehat{U}_{LT} = \widehat{U}_{RT} = \widehat{U}_{RS} = 1$，则由第一个边界条件得

$$(1-\mu_{LT}) + \alpha(1-\mu_{RT}) + \beta(1-\mu_{RS}) = 0, \tag{4.151}$$

将 (4.147) 代入，得

$$(1 - 1/\mu_{RT}) + \alpha(1-\mu_{RT}) + \beta(1+\mu_{RT}) = 0. \tag{4.152}$$

相似地，由第二个边界条件得

$$(1+1/\mu_{RT})\widehat{V}_{LT} + \alpha(1+\mu_{RT})\widehat{V}_{RT} + \beta(1-\mu_{RT})\widehat{V}_{RS} = 0. \tag{4.153}$$

从 (4.148) 算出 $\widehat{V}$ 并代入该方程，并利用恒等式

$$\frac{1/\mu}{1-(1/\mu)^2} = -\frac{\mu}{1-\mu^2}, \qquad \frac{-\mu}{1-(-\mu)^2} = -\frac{\mu}{1-\mu^2}, \tag{4.154}$$

[1] Matsuno, T.(1966), False reflection of waves at the boundary due to the use of finite differences, *J. Meteorol. Soc. Japan* **44**(2), 145-57.

再约去公因式 $\gamma\mu_{RT}/(1-\mu_{RT}^2)$ ，得

$$-(1+1/\mu_{RT})+\alpha(1+\mu_{RT})-\beta(1-\mu_{RT})=0. \tag{4.155}$$

最后，联立 (4.152) 和 (4.155) ，得到

$$\alpha=1/\mu_{RT}, \quad \beta=0, \tag{4.156}$$

这表明没有产生右行的伪解波型．另外也可看出 $j=\frac{1}{2}$ 时，向左右方向传播的真解波型具有相同的振幅，事实上，由 (4.152) 和 (4.147) 即可知 $\mu_{LT}^{1/2}=\alpha\mu_{RT}^{1/2}$.

对这个例子，我们给出了很多分析细节，部分是因为这种分析方法可以应用到其他问题以及其他格式中去 (例如对于 Lax-Wendroff 格式，边界条件 (4.150) 也很有效)，还因为这种分析方法与由边界条件选择失当所产生的不稳定性的一般分析方法属于同类 (参考书目见本章注记以及第 5.10 节) .

4.13 高维情形

从某些方面来讲，将本章的方法推广到二维和三维情形比第 2 、3 章中的抛物型方程要容易一些．原因主要是：对于双曲型方程我们通常不用也不必要使用隐式格式，这是因为显式格式的稳定性条件 $\Delta t=O(\Delta x)$ 并不苛刻，且为保证格式的精度也常令 Δt 和 Δx 大小差不多；另外对于重要的几类问题，比如数值天气预报，区域通常是周期性的，并不需处理很难的曲边边界．尽管如此，多维双曲型方程的理论发展得还远远不够充分，而且一维情形时某些特定格式的一些诱人的性质常常不能推广到二维或高维情形，而且，某些时候用隐式方法比较有利时，构造其产生的代数方程组的快速算法却很不容易 (参见第 7 章) .

一个典型的方程组是 (4.5) 和 (4.9) 的自然推广，可表示为

$$\frac{\partial \mathbf{u}}{\partial t}+\frac{\partial \mathbf{f}(\mathbf{u})}{\partial x}+\frac{\partial \mathbf{g}(\mathbf{u})}{\partial y}=0. \tag{4.157}$$

以及

$$\mathbf{u}_t+A(\mathbf{u})\mathbf{u}_x+B(\mathbf{u})\mathbf{u}_y=0, \tag{4.158}$$

其中雅可比矩阵 $A=\partial\mathbf{f}/\partial\mathbf{u}$ ， $B=\partial\mathbf{g}/\partial\mathbf{u}$ 由通量函数 $\mathbf{f}$ 和 $\mathbf{g}$ 导出．由 (4.90) 推广得到的非交错蛙跳格式可立即写出，且通过傅里叶分析，可给出标量情形的稳定性条件

$$\left|\frac{a\Delta t}{\Delta x}\right|+\left|\frac{b\Delta t}{\Delta y}\right|\leqslant 1. \tag{4.159}$$

交错蛙跳格式可以按不同的方式推广，且给出不同的稳定性条件 (参见习题 10) .

相似地，Lax-Wendroff 格式可简单地进行推广，其推广方式也会影响到稳定性条件.

例如，一个广泛使用的两步格式可写成如下形式

$$\mathbf{U}_{i+1/2,j+1/2}^{n+1/2} = \left[\mu_x\mu_y\mathbf{U} - \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\mu_y\delta_x\mathbf{F} + \frac{\Delta t}{\Delta y}\mu_x\delta_y\mathbf{G}\right)\right]_{i+1/2,j+1/2}^{n}, \tag{4.160}$$

$$\mathbf{U}_{i,j}^{n+1} = \mathbf{U}_{i,j}^{n} - \left[\frac{\Delta t}{\Delta x}\mu_y\delta_x\mathbf{F} + \frac{\Delta t}{\Delta y}\mu_x\delta_y\mathbf{G}\right]_{i,j}^{n+1/2} \tag{4.161}$$

为方便，其中引入了平均记号 $\mu_x U_{i,j} := \frac{1}{2}(U_{i+1/2,j} + U_{i-1/2,j})$. 对标量情形，其稳定性条件为

$$\left(\frac{a\Delta t}{\Delta x}\right)^2 + \left(\frac{b\Delta t}{\Delta y}\right)^2 \leqslant 1, \tag{4.162}$$

对于速度向量为 (a,b) 的波，这是 CFL 条件到高维情形的最自然的推广

如第 4.7 节所示，该格式是有限体积格式一个很好的例子；而且更新步 (4.161) 不仅对 Lax-Wendroff 格式，而且还对其他格式有价值，另外，可以将之推广到如图 1-1 和图 1-2 所示的四边形网格中去. 在这样的网格上， $\mathbf{U}_{i,j}^n$ 表示与质心相对应的单元平均，而 (4.161) 中的数值通量则需要在单元顶点上求值. 事实上，假定 α 表示绕单元逆时针旋转的循环坐标，仿 (4.66) 所做，先利用高斯散度定理，再对 (4.157) 中的通量项在 (i,j) 单元上积分. 且对边界上的积分应用梯形准则，得

$$\begin{aligned}|\Omega_{i,j}|(\mathbf{U}_{i,j}^{n+1} - \mathbf{U}_{i,j}^{n}) \quad + \quad & \frac{1}{2}\Delta t\sum_{\alpha}\{[\mathbf{F}_{\alpha+1} + \mathbf{F}_{\alpha}](y_{\alpha+1} - y_{\alpha}) \\ & - [\mathbf{G}_{\alpha+1} + \mathbf{G}_{\alpha}](x_{\alpha+1} - x_{\alpha})\}^{n+1/2} = 0.\end{aligned} \tag{4.163}$$

这里单元面积 $|\Omega_{i,j}|$（可通过在单元上对 $\operatorname{div}(x,0)$ 求积分得到）为

$$|\Omega_{i,j}| = \frac{1}{2}\sum_{\alpha}(x_{\alpha+1} + x_{\alpha})(y_{\alpha+1} - y_{\alpha}).$$

也请注意，对三角形区域也可推出相似的公式. 于是也可将之推广到由四边形和三角形组成的网格上.

上述过程中，较困难的一步是计算中间时间层 $n+\frac{1}{2}$ 顶点上的通量. 在 Lax-Wendroff 格式中使用了 (4.160)，我们可以通过以下方式将之推广，在由某顶点周围四个单元的中心组成的四边形上积分，得到相应的表达式. 但是因为 Lax-Wendroff 通量可能引起振荡，所以我们不能肯定这种方法会很有效，因此我们更愿意把 4.7 节中构造的 TVD 通量扩展到二维情形. 然而，对大多数双曲方程组来说，矩阵 A 和 B 不可交换，因此它们不能同时化为对角形，从而基于求解局部黎曼问题的迎风格式的高维形式很难构造.

因此，大多数高维的迎风格式使用的是边上某一内点的通量，而非顶点上的通量. 这时可利用双曲性质， α 和 β 为实数时，线性组合 $\alpha A + \beta B$ 的特征值总为实数，从而可

以使用法向上的通量. 这样就可以直接使用一维的算法, 并且只需要把更新步 (4.163) 中每边上所用的梯形准则改为中点准则或者高斯准则.

方法总是十分依赖于实际应用和考查的对象. 更多的细节讨论超出了本书的范围. 另外, 边界上也会产生相应的困难, 因为此时不能像一维那样简单地利用特征线的指向就可推出所需的边界条件.

文献注记与推荐读物

为了解有关双曲方程基本理论、特征线方法、黎曼不变量以及激波的更多内容, 可参考 Courant 和 Hilbert(1962) 以及 Courant 和 Friedrichs(1948) 等经典著作, 或者最近的几本书, 如 Carrier 和 Pearson(1976) 、 Smoller(1983) 以及 Evans(1998) . 为完整地了解非线性波和流体中的波, 可参考 Whitham(1974) 以及 Lighthill(1978).

一些课本结合守恒律的理论处理和激波问题的数值算法, 给出了更多的内容细节, 其中包括 LeVeque(1992, 2002) 和 Kreiss 和 Lorenz(1989) 所写的书. 对于双曲型方程初边值问题的适定性, 以及与数值边界条件和稳定性密切相关的内容, 后一本书也给出了权威的说明. 有关第 4.12 节中例子的分析方法的开创性的工作, 含在 Godunov 和 Ryabenkii 早期的论文 (1964) 中.

最近关于求解哈密顿系统的辛几何和辛积分算法的文章, 可参考 Hairer 等 (2002) 以及 Leimkuhler 和 Reich(2004) 所写的论文.

习　题

4.1　对方程 $u_t + au_x = 0$, 其中 $0 \leqslant x \leqslant 1$ 且 $a \equiv a(x) = x - \frac{1}{2}$, 画出特征线的简图. 在一致网格 $\{x_j = j\Delta x, j = 0, 1, \ldots, J\}$ 上建立迎风格式, 并推导出误差界, 注意此时不需要边界条件; 请分别考虑 J 是奇数和偶数的两种情况. 设初值为 $u(x, 0) = x(1-x)$, 画出解随时间发展的简图, 并通过估计截断误差中的项给出精确的误差界.

对 $a(x) = \frac{1}{2} - x$, 边界条件 $u(0, t) = u(1, t) = 0$ 的情形, 重复上面练习.

4.2　若 q 可展开成 p 的幂形式如下

$$q \sim c_1 p + c_2 p^2 + c_3 p^3 + c_4 p^4 + \cdots,$$

证明

$$\tan^{-1} q \sim c_1 p + c_2 p^2 + (c_3 - \frac{1}{3} c_1^3) p^3 + (c_4 - c_1^2 c_2) p^4 + \cdots,$$

即第 4.4 节中的引理 4.1．利用以上结果，推导出用下列格式逼近 $u_t + au_x = 0$ 时，相位展开式的主项：

迎风格式	$-\nu\xi + \frac{1}{6}\nu(1-\nu)(1-2\nu)\xi^3$;
Lax-Wendroff 格式	$-\nu\xi + \frac{1}{6}\nu(1-\nu^2)\xi^3$;
盒式格式	$-\nu\xi - \frac{1}{12}\nu(1-\nu^2)\xi^3$;
蛙跳格式	$-\nu\xi + \frac{1}{6}\nu(1-\nu^2)\xi^3$;

其中 $\nu = a\Delta t/\Delta x$， $\xi = k\Delta x$．

4.3　验证由方程

$$u = f(x - ut)$$

隐式定义的函数 $u(x,t)$ 是下面问题的解

$$u_t + uu_x = 0, \qquad u(x,0) = f(x),$$

并且说明 $u(x,t)$ 在直线 $x - x_0 = tf(x_0)$ 上等于常数 $f(x_0)$．

证明当 ϵ 很小时，过点 $(x_0, 0)$ 和 $(x_0+\epsilon, 0)$ 的直线交于一点，且 $\xi \to 0$ 时，交点的极限是 $(x_0 - f(x_0)/f'(x_0), -1/f'(x_0))$．并证明若对任意的 x 有 $f'(x) \geqslant 0$，则对任意的正的 t，解都是单值的．更一般地，若 $f'(x)$ 有时取负值，则存在 $t_c = -1/M$，其中 M 是 $f'(x)$ 取得的最大负值，满足解在 $0 \leqslant t \leqslant t_c$ 时是单值的．

证明对于函数 $f(x) = \exp[-10(4x-1)^2]$，临界值为 $t_c = \exp(\frac{1}{2})/8\sqrt{5}$，约等于 0.092．[请对照图 4-10.]

4.4　设 a 为正常数，考虑求解方程 $u_t + au_x = 0$ 的格式

$$U_j^{n+1} = c_{-1}U_{j-1}^n + c_0U_j^n + c_1U_{j+1}^n,$$

试确定系数 c_0, c_1, c_{-1}，使其对方程的解有与 $u(x_j, t_{n+1})$ 的泰勒展开式有尽可能多的项相吻合．验证所得的格式就是 Lax-Wendroff 格式．

按同样的方式确定下面格式的系数

$$U_j^{n+1} = d_{-2}U_{j-2}^n + d_{-1}U_{j-1}^n + d_0U_j^n.$$

验证所得的系数 d 相对于 Lax-Wendroff 格式的系数，仅仅是将 c 中的 ν 换成 $\nu - 1$．通过把微分方程作变量替换 $\xi = x - \lambda t$，解释其原因，其中 $\lambda = \Delta x/\Delta t$．并由此求出该格式的稳定性条件．

4.5　对标量情形的守恒律 $u_t + f(u)_x = 0$，其中 $f(u)$ 仅是 u 的函数， Lax-Wendroff 格式可写为

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}P_j^n + \left(\frac{\Delta t}{\Delta x}\right)^2 Q_j^n,$$

其中

$$P_j^n = \frac{1}{2}\left[f(U_{j+1}^n) - f(U_{j-1}^n)\right],$$

$$Q_j^n = \frac{1}{2}\left[A_{j+1/2}^n(f(U_{j+1}^n) - f(U_j^n)) - A_{j-1/2}^n(f(U_j^n) - f(U_{j-1}^n))\right],$$

且

$$A_{j+1/2}^n = f'(\frac{1}{2}U_{j+1}^n + \frac{1}{2}U_j^n).$$

在点 (x_j, t_n) 处做泰勒展开，验证 P 和 Q 的展开式中分别只含 Δx 的奇次幂项和偶次幂项. 验证格式截断误差的主项是

$$T_j^n = \frac{1}{6}(\Delta t)^2 u_{ttt} + (\Delta x)^2\left[\frac{1}{6}u_{xxx}f' + \frac{1}{2}u_x u_{xx} f'' + \frac{1}{6}u_x^3 f'''\right].$$

4.6 对线性对流方程 $u_t + au_x = 0$，其中 a 是正常数，均匀网格上的一个广义的迎风格式定义如下

$$U_j^{n+1} = (1-\theta)U_k^n + \theta U_{k-1}^n$$

其中 $x_k - \theta\Delta x = x_j - a\Delta t$ 并且 $0 \leqslant \theta < 1$. 验证 CFL 条件不需要对 Δt 做任何限制，并且冯诺伊曼分析表明格式是无条件稳定的. 求出格式的截断误差，并说明其随 Δt 的增长如何变化.

4.7 微分方程

$$u_{xx} - (1+4x)^2 u_{tt} = 0$$

定义在区域 $0 < x < 1,\ t > 0$ 上，其初边值条件是

$$u(x,0) = x^2,\ \ u_t(x,0) = 0,\quad u_x(0,t) = 0,\ \ u(1,t) = 1.$$

构造出求解该方程的显式中心差分格式. 说明边界条件在数值格式中是如何应用的. 写出微分方程的特征线，并利用 CFL 条件给出一个稳定性约束.

4.8 线性化的一维形式的可压缩等熵流方程是

$$\rho_t + q\rho_x + w_x = 0,$$
$$w_t + qw_x + a^2\rho_x = 0,$$

其中 a 和 q 是正常数. 证明用中心差分逼近 x 的导数的显式格式恒不稳定. 在此基础上加上由 Lax-wendroff 格式产生的附加项，导出一个条件稳定的差分格式，并找出其稳定性条件.

4.9 对线性对流方程 $u_t + au_x = 0$，利用差分 $U_j^{n+1} - U_j^n$ 和 $U_{j+1}^n - U_{j-1}^{n+1}$ 构造一个“斜角导数”(angled-derivative) 差分格式. 利用傅里叶分析研究格式的精度和稳定性. 对于 $a > 0$，确定在区域 $0 \leqslant x \leqslant 1$ 上求解此问题的边界条件，并且考虑在每一时间步上解的发展情况.

4.10 对于二维波动方程

$$\rho_t + u_x + v_y = 0, \quad u_t + c^2\rho_x = 0, \quad v_t + c^2\rho_y = 0,$$

在步长为 h 的正方形网格上，构造一个交错型蛙跳格式，使得 ρ 定义在点 $x = rh, y = sh$ 上，(u, v) 定义在点 $x = (r + \frac{1}{2})h, y = (s + \frac{1}{2})h$ 上，并找出其稳定性条件.

另外构造一个交错型蛙跳格式，使得 ρ 定义在原来的点上， u 定义在点 $x = (r + \frac{1}{2})h$, $y = sh$ 上， v 定义在 $x = rh, y = (s + \frac{1}{2})h$ 上. 找出其稳定性条件，并与前一格式比较，证明其中一个比另一个有更好的稳定性.

4.11 证明均匀网格上的 Engquist-Osher 格式 (4.77) 当其稳定时是 TVD 的.

证明对于非均匀网格上的 Roe 格式 (4.74)，当其时间步满足

$$-\Delta x_j \leqslant A^n_{j+1/2}\Delta t \leqslant \Delta x_{j+1},$$

时，也就是过单元边界 $x_{j+1/2}$ 的特征线在一个时间步中不超出其旁边的单元时，格式是 TVD 的.

第 5 章　相容性、收敛性和稳定性

5.1　问题的定义

本章将汇集前面各章引入的定义，并将其严格化．这样，我们就能够叙述并证明关键的 *Lax* 等价定理(Lax Equivalence Theorem) 的主要部分．为简单起见，我们并不追求普适性，不过我们的定义和论证与更一般的论述是一致的．

对所要考虑的问题，我们做出以下假设：

- 区域 Ω 是一维、二维、三维或更高维空间中的固定的有界开区域，其坐标系可能是笛卡儿坐标 $(x, y, \ldots)$，圆柱坐标，球坐标，等等；
- 区域 Ω 的边界为 $\partial\Omega$；
- 所要找的解是定义在 $\Omega \times [0,\ t_F]$ 上的关于空间变量和时间 t 的函数 u．该函数可以是向量值的，因此所讨论的结果不仅适用于单个方程，也适用于微分方程组；
- 算子 $L(\cdot)$ 含有 u 的关于空间变量的偏导数， L 不显含时间 t，在大部分情况下我们都假设 L 是线性算子，不过我们将尽可能使所给出的定义和所叙述的结果可以方便地推广到非线性算子；
- 在区域 Ω 的部分或全部边界上边界条件会给出了 $g(u)$ 的值，其中 $g(\cdot)$ 是可能含有空间变量偏导数的算子；
- 初始条件在 Ω 上给出了 u 在 $t = 0$ 时的值．

由此，我们将所考虑的问题写成一般形式

$$\frac{\partial u}{\partial t} = L(u), \qquad 在\ \Omega \times (0,\ t_F]\ 内, \tag{5.1a}$$

$$g(u) = g_0, \qquad 在\ \partial\Omega_1 \subset \partial\Omega\ 上, \tag{5.1b}$$

$$u = u^0, \qquad 在\ \Omega\ 上，当\ t = 0\ 时. \tag{5.1c}$$

我们将始终假定 (5.1) 定义了一个*适定问题* (well-posed problem)，其意义将在稍后阐明，粗略地说就是，问题的解总是存在的且连续地依赖于所给数据．

5.2 有限差分的网格与范数

设差分逼近定义在固定的网格上，设在不同网格点和时间步上，时间步长都是相同的常数 Δt. 为简单起见，通常设区域 Ω 上的网格有一致的步长，在笛卡儿坐标时步长为 $\Delta x, \Delta y, \ldots$; 在极坐标时步长为 $\Delta r, \Delta\theta, \ldots$. 每个网格点上的函数值记为 U_j^n, 在二维或更高维空间，下标 j 为多重指标，作为简洁的记号来代替 $U_{j,k}^n, U_{j,k,l}^n$ 等. 设用一个固定的正则差分格式求解 U_j^n, 其中下标 j 属于集合 J_Ω, 则只有该集合中的指标会出现在范数的定义中. 通常该集合恰好包含网格的内点；当然必要时，如在有弯曲的边界、含导数的边界条件等情况下需要外推至虚拟的外部节点，正则的差分格式可以推广到所有需要的点上，例如见 3.4 节及 (3.35) 的使用. 还有另外一些需要考虑的情况，正则有限差分算子作用到对称的边界点上，见下面的 6.5 节，或作用在有周期边界条件的边界点上，这时这些边界点也应包含在 J_Ω 中. 在第 n 个时间层上，U 在所有这些点上的取值记为 U^n:

$$U^n := \{U_j^n,\ j \in J_\Omega\}. \tag{5.2}$$

为简单起见，我们仅考虑含两个时间层的格式，对于 1- 步格式这意味着对每个 U_j^n, 如果看作向量，则与 u 的维数相同. 不过，这并不排除多层格式，如 4.9 节中见过的蛙跳格式，为此只需扩展 U_j^n 相对于 u 的维数即可. 例如，若格式涉及三个时间层，即 U^{n+1} 用 U^n 和 U^{n-1} 算出，则可定义维数为原来两倍的新向量 $\tilde{U}^n$, 其元素即为 U^n 和 U^{n-1} 中的元素.

为比较 U 与 u, 需要引入适用于两者，尤其是两者之差的范数. 为此，首先记函数 $u(x,t)$ 的网格值为 u_j^n, 这通常就是网格节点值 $u(x_j, t_n)$. 我们希望证明 U 的网格值收敛于 u 的网格值. 因此，与网格点值 U_j^n 类似地定义

$$u^n := \{u_j^n,\ j \in J_\Omega\}. \tag{5.3}$$

我们仅考虑两种范数. 首先定义**最大模范数** (maximum norm):

$$\|U^n\|_\infty := \max\{|U_j^n|,\ j \in J_\Omega\}. \tag{5.4}$$

如果将 u 看作是 t_n 固定时以 x 为变量的函数，则 u^n 的最大模范数近似等于 u 的通常意义下的极大模范数 (supremum norm)，但一般并不严格相等，两个范数值仅当函数 $|u(x,t_n)|$ 的最大值恰在网格节点上取到时才会相等.

其次用到的是一种离散的 l_2 范数，这种范数可以看作是 L_2 积分范数的近似. 为了定义该范数，我们首先对每个网格内点引入测度为 V_j 的“控制体 (control volume)”，这些控制体形成了一族互不重叠的单元，并近似地覆盖了 Ω. 通常，如图 5-1 所示，网格点 x_j 处于相应的控制体的中心，亦见 4.7 节有限体积法；但这并非必要，所需的只是网格点与控制体间存在一一对应. 对于三维笛卡儿几何，$V_j = \Delta x\Delta y\Delta z$; 而对于三维圆柱几

何， $V_j = r_j \Delta\theta \Delta r \Delta z$, 等等. 因此，我们定义

$$\|U^n\|_2 := \left\{ \sum_{j \in J_\Omega} V_j |U_j^n|^2 \right\}^{1/2} . \tag{5.5}$$

对于靠近边界的网格点，相应的控制体可能需要修正使其完全落在 Ω 内，但也可能不需要这么做. 无论如何， (5.5) 中的和式显然是一个积分近似，而且 $\|u^n\|_2$ 就近似等于 u 在时刻 t_n 的 L_2 积分范数

$$\|u(\cdot, t_n)\|_2 := \left[\int_\Omega |u(x, t_n)|^2 \,\mathrm{d}V \right]^{1/2}, \tag{5.6}$$

但两者一般并不严格相等. 然而，若令 u_j^n 等于 $u(x, t_n)$ 的平方在控制体上的积分平均值开方，则显然两者恰好严格相等；我们在 2.14 节模拟热量守恒性质时，曾见到过可做类似解释的量. 对于单个微分方程，符号 U_j^n 的意义不言自明；如果是微分方程组， U_j^n 是向量，而 $|U_j^n|$ 是该向量的范数. 具体选用那种向量范数对以后的分析无关紧要，当然前后的选择必须相互一致.

我们或许应该注意，这个一般的框架并没有包含某些实际当中常用的方法，其中许多与自适应方法有关：下一时间步的步长选取依赖于当前步的误差估计；与迎风格式 (4.20) 相同，根据解的性态选择向后或向前差分；或者局部加密网格，例如，在梯度大的地方. 其中一些做法要求在分析方法上做出重大的改变，而另一些，如非一致的网格加密，却不会带来太多的困难.

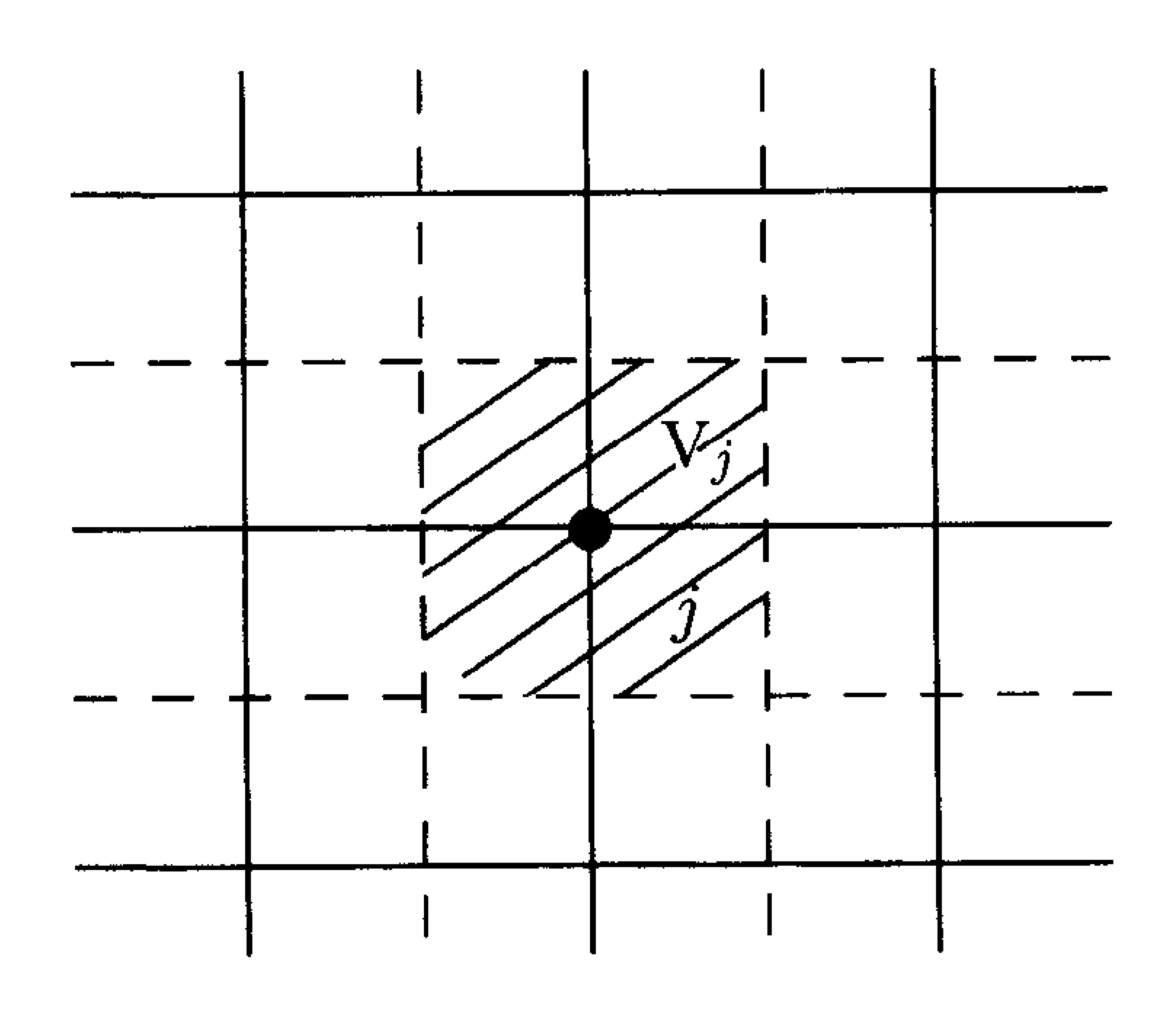


图 5-1 控制体的定义

5.3 有限差分逼近

我们要考虑的差分格式的一般形式为

$$B_1U^{n+1}=B_0U^n+F^n. \tag{5.7}$$

如符号所暗示，差分算子 B_0,B_1 与 n 无关，与此相对应的假设是，$L(\cdot)$ 不显式地依赖于 t. 然而，虽然基于固定的差分算子，B_0, B_1 却可能依赖于所作用的点. 因此，对每一点 $j\in J_\Omega$，线性差分算子 B 将被写成 J_Ω 中的紧邻点的线性组合：

$$(BU^n)_j=\sum_{k\in J_\Omega}b_{j,k}U_k^n \qquad \forall j\in J_\Omega, \tag{5.8}$$

而非线性算子就会包括 U_j^n 的非线性组合. 记符号 $b_{j,k}$ 反映了这些系数出于两个原因不仅依赖于 k 而且也会依赖于 j，原因之一是，这样可以包括 $L(\cdot)$ 含有空间变量系数的情况，而对常系数问题，$b_{j,k}$ 通常只依赖于差值 $k-j$. 另一个原因是，尽管对所有远离边界的点，(5.8) 中所用到的邻近点的相对位置都是一样的；但在边界点上，由于 (5.8) 已经结合了数值边界条件因而不再含有 U 在 J_Ω 之外的节点上的值. 所以，(5.7) 中的数据项 F^n 不仅包含由微分算子 $L(u)$ 的非齐次项产生的数据，也包含非齐次边界的数据.

设 B_1 总可以表示为形如 (5.8) 的线性算子，并因而可以用方阵来表示. 将有关理论推广到非线性问题只需令 B_0 为非线性的，而不必要求 B_1 是非线性的；但要包含盒式格式这样的算法，则要求 B_1 也是非线性的.

进一步假设 B_1 是可逆的，即其表示矩阵是非奇异的. 因此 (5.7) 式可写为

$$U^{n+1}=B_1^{-1}[B_0U^n+F^n]. \tag{5.9}$$

假定 (5.7) 做了适当的尺度调整使其取极限时，形式上表示了相应的微分方程，因而有 $B_1=O(1/\Delta t)$. 于是，当网格步长 $\Delta t,\Delta x,\ldots$ 以某种或需满足相容性条件的方式加密时，有

$$B_1u^{n+1}-[B_0u^n+F^n]\longrightarrow\frac{\partial u}{\partial t}-L(u). \tag{5.10}$$

以第 2 章中讨论的一维扩散方程的 θ-方法 (2.75) 为例，在远离边界时有

$$B_1=\frac{1}{\Delta t}-\theta\frac{\delta_x^2}{(\Delta x)^2},\quad B_0=\frac{1}{\Delta t}+(1-\theta)\frac{\delta_x^2}{(\Delta x)^2}. \tag{5.11}$$

结合前两个条件再进一步假设 B_1 是一致良态的，即尽管 B_1^{-1} 表示矩阵的阶数当 $\Delta t\to 0$ 时趋向于无穷，但对某个用来做分析的范数存在常数 K, 使得

$$\|B_1^{-1}\|\leqslant K_1\Delta t. \tag{5.12}$$

以 (5.11) 为例，取最大模，则不难推出 $K_1=1$. 过程如下：记 $\mu=\Delta t/(\Delta x)^2$, 方程

$$B_1U=F,\qquad U=B_1^{-1}F \tag{5.13}$$

在远离边界的点处可展开为

$$-\mu\theta U_{j-1} + (1+2\mu\theta)U_j - \mu\theta U_{j+1} = \Delta t F_j,$$

即

$$(1+2\mu\theta)U_j = \Delta t F_j + \mu\theta(U_{j-1} + U_{j+1}). \tag{5.14}$$

在狄利克雷边界条件下， $B_1 U = F$ 在 $j=1$ 和 $j = J-1$ 处仅涉及 $J_\Omega \equiv \{1, 2, \ldots, J-1\}$ 中的两个点，即可分别展开为

$$(1+2\mu\theta)U_1 - \mu\theta U_2 = \Delta t F_1 \quad \text{和} \quad -\mu\theta U_{J-2} + (1+2\mu\theta)U_{J-1} = \Delta t F_{J-1}.$$

因此对所有的 $j \in J_\Omega$，都有

$$(1+2\mu\theta)|U_j| \leqslant \Delta t\|F\|_\infty + 2\mu\theta\|U\|_\infty,$$

即

$$(1+2\mu\theta)\|U\|_\infty \leqslant \Delta t\|F\|_\infty + 2\mu\theta\|U\|_\infty. \tag{5.15}$$

由此即得所证.

5.4 相容性、精度的阶和收敛性

所有的极限运算和渐近结果都是 (有时是隐含地) 指关于一个网格加密路径或一族网格加密路径. 即如 (2.47), 选择网格参数 Δt,Δx,Δy 等的一个收敛于零的序列，并注意它们之间可能需要满足的不等式约束关系. 为简单起见，我们仅用单一的参数 h 来表征整个空间的离散化，这个参数也许只不过是网格步长 Δx,Δy,... 中最大的一个，也或许还需要根据各坐标方向上的特征速度做尺度变换； h 也可能是网格点周围所有控制体的直径中最大的一个. 然后沿着某个设计好的网格加密路径取极限，记为 “ $\Delta t(h) \to 0$ ”，也有时简记为 $\Delta t \to 0$ 或 $h \to 0$. 我们总要求步长 Δt 趋于零，不过稳定性或相容性可能会要求其收敛于零的速率决定于 h, 例如，对于抛物型问题，典型的是 $\Delta t = O(h^2)$, 而对双曲型问题则是 $\Delta t = O(h)$.

截断误差(truncation error) 用精确解 u 定义如下

$$T^n := B_1 u^{n+1} - [B_0 u^n + F^n], \tag{5.16}$$

而差分格式 (5.7) 关于问题 (5.1a)~(5.1c) 的*相容性* (consistency) 则定义为,对所有 (5.1a)~(5.1c) 的充分光滑的解 u 均有

$$T_j^n \longrightarrow 0, \quad \text{当} \ \Delta t(h) \to 0, \ \forall j \in J_\Omega. \tag{5.17}$$

注意这里已经包含了边界条件的相容性，因为在 B_0 和 B_1 的定义中已消去了边界上的 U 值.

设对充分光滑的解 u, p 和 q 是使

$$|T_j^n| \leqslant C[(\Delta t)^p + h^q], \quad 当\ \Delta t(h) \to 0,\ \forall j \in J_\Omega. \tag{5.18}$$

成立的最大整数，则称相应格式的精度关于 Δt 的阶为 p, 关于 h 的阶为 q, 或关于 Δt 的精度的阶 (oder of accuracy) 为 p, 关于 h 的精度的阶为 q.

另一方面，收敛性(convergence) 的定义则关系到使得问题 (5.1a)~(5.1c) 适定的所有初始数据及其他数据，其意义将在下节定义. (5.7) 称作是在范数 $\|\cdot\|$ 下为 (5.1a)~(5.1c) 提供了收敛的逼近(convergent approximation), 如果对每个在范数 $\|\cdot\|$ 的意义下，使 (5.1a)~(5.1c) 适定的 u^0 都有

$$\|U^n - u^n\| \longrightarrow 0, \quad 当\ \Delta t(h) \to 0,\ n\Delta t \to t \in (0,\ t_F], \tag{5.19}$$

这里所用的范数是指 (5.4) 或 (5.5). 从实用的观点看，这种做法的优越性在于直接吸纳了舍入误差的影响，如果稳定性的定义仅适用于充分光滑的数据，则舍入误差的影响必须另做分析.

5.5 稳定性与 Lax 等价定理

上节的定义 (5.16)~(5.19) 并不仅局限于线性问题，事实上它们具有相当的一般性. 但是在本节 (及本章以后的大部分) 中，我们所能考虑的仅限于线性问题. 设 V^n 和 W^n 是 (5.7) 或 (5.9) 的具有同样非齐次项 F^n 的解，但它们的初值分别为 V^0 和 W^0, 我们称该算法关于范数 $\|\cdot\|$ 和给定的网格加密路径为稳定的(stable), 如果存在常数 K, 使得

$$\|V^n - W^n\| \leqslant K\|V^0 - W^0\|, \qquad n\Delta t \leqslant t_F, \tag{5.20}$$

其中，常数 K 有与 V^0, W^0 和网格加密路径中的 $\Delta t(h)$ 无关的一致的界. (在非线性情形，通常必须对初值加以限制.) 假设 V^n 和 W^n 有相同的数据项 F^n, 这是对问题的简化，主要是因为我们不想过多地考虑边界条件对于本章和前面各章中所研究的发展型问题的影响 (这也是我们应用傅里叶分析研究稳定性时可以令边界条件为零的原因). 边界条件的影响在第 6 章讨论椭圆型问题时将予以更多的讨论. 也请注意，如同第 2 章 2.11 节所做，通过应用 Duhamel 原理，我们也能够考虑内点上数据的影响.

由于考虑的是线性问题， (5.20) 可以写为

$$\|(B_1^{-1}B_0)^n\| \leqslant K, \qquad n\Delta t \leqslant t_F. \tag{5.21}$$

注意对于隐式格式，证明 (5.12) 式成立是建立 (5.21) 式的重要一步；例如，考虑线性对流问题的盒式格式，并考虑由于在某一侧边界上给定边界条件而带来的影响时.

现在可以正式地定义适定性 (well-posedness) 了. 问题 (5.1) 称为关于范数 $\|\cdot\|$ 是适定的 (well-posed), 如果对所有充分小的 h, 可以证明： (i) 对任意初始值 u^0, 只要 $\|u^0\|$ 关于 h

一致有界，则解存在； (ii) 存在常数 K'，使得对任意一对解 v 和 w，都有

$$\|v^n - w^n\| \leqslant K'\|v^0 - w^0\|, \qquad t_n \leqslant t_F. \tag{5.22}$$

该定义与通常的定义的区别在于这里使用了离散范数. 不过我们所选取的两个离散范数当 $h \to 0$ 时与相应的函数范数等价，前提是 u 的函数范数有界，且 u_j^n 有适当的定义. 不论是使用离散范数还是函数范数，这种适定性定义的一个重要性质是，虽然 (5.1a) 的古典解 u 必须充分光滑以使得相应的导数存在，但是假设有一个使得光滑解存在的初始值序列，且该序列依 $\|\cdot\|$ 范数关于 h 一致地收敛于一任意的初始值 u^0, 则由 (5.22) 我们可以定义对应于该初始值的 *广义解* (generalised solution), 其在任意时刻 t^n 处的取值等于相应的光滑初始值的解序列的极限值. 因此，在建立问题的适定性时只需对一个光滑的稠密集证明其解的存在性，当然这里所定义的稠密性必须是关于 h 是一致的.

显然，微分方程问题适定性的定义与由 (5.20) 给出的离散问题稳定性的定义之间关系十分密切. 这个由 Lax 于 1953 年引入的稳定性定义，使他得以导出以下极为重要的定理.

定理 5.1. (Lax 等价定理) *对一个适定的线性发展问题在 (5.12) 意义下一致可解的相容的差分逼近格式，格式的稳定性是其收敛性的充分必要条件.*

证明. (充分性). 将 (5.7) 与 (5.16) 相减得

$$B_1(U^{n+1} - u^{n+1}) = B_0(U^n - u^n) - T^n,$$

即

$$U^{n+1} - u^{n+1} = (B_1^{-1}B_0)(U^n - u^n) - B_1^{-1}T^n. \tag{5.23}$$

设取 $U^0 = u^0$, 则有

$$\begin{aligned} U^n - u^n \;\; = \;\; & -[B_1^{-1}T^{n-1} + (B_1^{-1}B_0)B_1^{-1}T^{n-2} + \cdots \\ & +(B_1^{-1}B_0)^{n-1}B_1^{-1}T^0]. \end{aligned} \tag{5.24}$$

在定理的应用中， (5.12) 式和 (5.21) 式用的是同一个范数，且该范数也要用于 (5.19) 的推导，结合两式便得

$$\|(B_1^{-1}B_0)^m B_1^{-1}\| \leqslant KK_1\Delta t, \tag{5.25}$$

由 (5.24) 给出

$$\|U^n - u^n\| \leqslant KK_1\Delta t \sum_{m=0}^{n-1} \|T^m\|.$$

又若 u 充分光滑因而相容性 (5.17) 成立，则立即得到在 (5.19) 意义下的收敛性. 解不够光滑时的收敛性则得证于适定性和稳定性的假设：一般初始数据可用光滑解的数据任意

逼近，而由微分问题的适定性和离散问题的稳定性 (5.19) 对应不同初始值的解之间差异的增长是有界的.

用泛函分析中的共鸣定理便可证明稳定性是收敛性的必要条件，不过要求连续问题和离散问题放在同一个 Banach 空间的框架中考虑；而这正是基于离散范数的简化做法的不足之处，因此考虑应用该定理已超出本书范围，感兴趣的读者可以在 Richtmyer 和 Morton (1967), pp. 34-36, 46 中找到相应的证明. ■

由上述定理，对于任何容易证明相容性的格式，真正需要关心的只是确定稳定性条件，因此只需限于研究离散问题. 如我们所见，通常相容性对任何 $\Delta t \to 0, h \to 0$ 的序列都成立，但有几种情况必须加以注意. 例如，一维热传导方程的 Dufort-Frankel 格式

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} = \frac{U_{j+1}^n - U_j^{n+1} - U_j^{n-1} + U_{j-1}^n}{(\Delta x)^2}, \tag{5.26}$$

该格式的优点是：它是显式格式但同时是无条件稳定的，这与我们到目前为止的经验相违背. 不过，其截断误差为

$$T = (u_t - u_{xx}) + (\Delta t/\Delta x)^2 u_{tt} + O((\Delta t)^2 + (\Delta x)^2) + ((\Delta t)^2/\Delta x^2). \tag{5.27}$$

因此，该格式仅当 $\Delta t = O(\Delta x)$ 时与热传导方程相容，且仅当 $\Delta t = O((\Delta x)^2)$ 时有一阶精度. 其结果是，相容性条件而非稳定性条件决定了能保证收敛性的加密路径.

该例子突显出在 Lax 等价定理中，不仅隐含了稳定性和收敛性定义中范数的选取，而且隐含了加密路径的选取.

5.6　稳定性条件的计算

由于涉及的是线性问题，若 (5.20) 中的 V^n 和 W^n 都是差分方程 (5.7) 的解，则差 $V^n - W^n$ 是满足齐次边界条件的齐次差分方程的解. 因此，证明稳定性等价于证明 [1]：

$$B_1 U^n = B_0 U^{n-1} \quad \text{和} \quad n\Delta t \leqslant t_F \quad \Longrightarrow \quad \|U^n\| \leqslant K \|U^0\|, \tag{5.28}$$

即等价于证明 (5.21). 常数 K 一般依赖于时间段 t_F 且允许某种形式的指数增长 (例如方程 $u_t = u_x + u$ 的解可能会出现指数增长的情况). 对于简单的问题，情况经常是，要么 $K = 1$, 这时误差没有增长，因而格式是稳定的；要么对某种波型即使当 $\Delta t \to 0$ 时，仍有 $U^n \sim \lambda^n U^0$, 其中 $|\lambda| > 1$, 因而格式是不稳定的.

因此，我们在 2.6 节及其他地方证明了最大值原理，同时也就证明了最大模范数意义下的稳定性：严格地讲，还必须证明最小值原理，即不仅有

$$U_j^{n+1} \leqslant \max_k U_k^n \leqslant \|U^n\|_\infty, \tag{5.29}$$

[1] 原书中是 $B_1 U^{n+1} = B_0 U^n$, 但这样应要求 $(n+1)\Delta t \leqslant t_F$, 相应的结论应为 $\|U^{n+1}\| \leqslant K\|U^0\|$.

还应有

$$U_j^{n+1} \geqslant \min_k U_k^n \geqslant -\|U^n\|_\infty, \tag{5.30}$$

因而可以推出

$$\|U^{n+1}\|_\infty \leqslant \|U^n\|_\infty. \tag{5.31}$$

以上证明稳定性的方法对于抛物型问题十分自然，因为对于这类问题最大值 (或最小值) 原理是差分格式所具有的非常有吸引力的恰当的且强于稳定性的属性. 同样的方法，如前面的例子所示，可以处理带有变系数和混合边界条件的相当一般的线性问题：在每个例子中都得以推出使格式满足最大值原理时， Δt 所应满足的简单的代数条件，因此这些条件满足时就得到了算法在最大模范数意义下的稳定性. 在许多情况下，对相应的带有常系数和周期边界条件的问题，通过傅里叶分析的方法可以证明，若这些条件不满足，则会导致 l_2 范数意义下的不稳定性 (见后面的讨论). 然而，在另外一些情况下，例如对于 θ-方法，两种方法给出的稳定性条件是有差距的. Fritz John 在他 1952 年发表的一篇有影响的论文 [1] 中证明，对于更广泛的一类抛物型问题和相应的差分逼近格式，只要格式满足由局部傅里叶分析得到的冯诺伊曼条件，则格式在最大模范数的意义下也是稳定的. 因此，一些格式虽不满足最大值原理但实际上仍然在最大模范数意义下是稳定的 (关于这方面更多的评注见下面的 5.7 节).

再者，对双曲型问题，最大值原理一般不成立，甚至提法也不适宜. 前面我们曾提到过，一阶格式 (4.20) 当 $0 \leqslant \nu \leqslant 1$ 时满足最大值原理，因而按最大模范数是稳定的；但我们可以证明二阶格式则不可能满足最大值原理. 例如，考虑表示为 (4.36) 的 Lax-Wendroff 格式. 若其满足最大值原理，则对任意取非正值的 U^n, 不可能出现 $U^{n+1} > 0$ 的情况，但当 $0 < \nu < 1$ 时，令 $U_{j-1}^n = U_j^n = 0$ 及 $U_{j+1}^n = -1$, 则有 $U_j^{n+1} > 0$, 当然这并不表明格式实际上按最大模范数是不稳定的，只不过我们不能用这种方法来证明其稳定性.

由于这一原因，同时也由于双曲型微分方程通常更多地在 L_2 范数下而不是在极大模范数下是适定的，对于双曲型问题我们不得不采用较弱的 l_2 范数 (5.5) 来证明稳定性，这给出了较弱的结果，因为在进行计算的有界区域上有

$$\left[\min_{j\in J_\Omega} V_j\right]^{1/2} \|U\|_\infty \leqslant \|U\|_2 \leqslant \left[\sum_{j\in J_\Omega} V_j\right]^{1/2} \|U\|_\infty, \tag{5.32}$$

其中 V_j 是第 j 个控制体的测度，右端的系数是有界的，而当网格加密时左端的系数趋向于零. 我们当然更希望通过稳定性分析得到最大模范数下的误差界，但如果只有 l_2 稳定

[1] John, F. (1952). On the integration of parabolic equations by difference methods, *Comm. Pure Appl. Math.* **5**, 155.

性，则只能得到 l_2 范数下 $\|E\|_2$ 的误差界，再由 (5.32) 得到很差的最大模范数下 $\|E\|_\infty$ 的误差界.

然而，由于有 Parseval 关系式， l_2 范数恰恰适合于傅里叶分析. 设在一个正则化的区域 $[-\pi,\ \pi]^d$ 上有一个一致的 (笛卡儿) 网格，网格尺度为 $\Delta x_1 = \Delta x_2 = \cdots = \Delta x_d = \pi/J$，并设边界条件为周期的，则在网格上可以彼此区别开来的傅里叶波型对应于波数，用向量 $\mathbf{k}$ 表示，它的分量为

$$k = 0, \pm 1, \pm 2, \ldots, \pm J, \tag{5.33}$$

其中事实上最后两项所代表的 $k\Delta x = \pm\pi$ 是区别不开的. 因此，该网格上的周期函数可展开为

$$U(\mathbf{x}_j) = \frac{1}{(2\pi)^{d/2}} \sum\nolimits'_{(\mathbf{k})} \widehat{U}(\mathbf{k}) \mathrm{e}^{\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}_j}, \tag{5.34}$$

其中求和号上面的一撇表示求和时，所有相应于 $k_s = \pm J$ 项的权都要除以 2, 而网格点用向量 $\mathbf{x}_j$ 表示. 这个离散傅里叶展开有一个逆变换，即离散傅里叶变换

$$\widehat{U}(\mathbf{k}) = \frac{1}{(2\pi)^{d/2}} \sum\nolimits'_{(j)} (\Delta x)^d U(\mathbf{x}_j) \mathrm{e}^{-\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}_j}, \tag{5.35}$$

其中 j 的每个分量从 $-J$ 到 J 取值，且也在网格的周期边界点处将相应的权除以 2, 这样所有项的权都等于 (5.5) 定义的 V_j.

引理 5.1. 分量由 (5.33) 给出的傅里叶波型 $(2\pi)^{-d/2}\mathrm{e}^{\mathrm{i}\,\mathbf{k}\cdot\mathbf{x}_j}$ 关于 (5.35) 中使用的 l_2 内积，即

$$\langle U,\ W\rangle_2 := (\Delta x)^d \sum\nolimits'_{(j)} U_j \overline{W}_j \tag{5.36}$$

是正交的.

证明. 只需考虑 $d = 1$ 的情形. 首先来建立基本的三角恒等式

$$\frac{1}{2}\mathrm{e}^{-\mathrm{i}\,J\theta} + \mathrm{e}^{-\mathrm{i}\,(J-1)\theta} + \cdots + \mathrm{e}^{\mathrm{i}\,(J-1)\theta} + \frac{1}{2}\mathrm{e}^{\mathrm{i}\,J\theta} = \sin J\theta \cot \frac{1}{2}\theta. \tag{5.37}$$

在和式

$$1 + \mathrm{e}^{\mathrm{i}\,\theta} + \mathrm{e}^{\mathrm{i}\,2\theta} + \cdots + \mathrm{e}^{\mathrm{i}\,(J-1)\theta} = (\mathrm{e}^{\mathrm{i}\,J\theta} - 1)/(\mathrm{e}^{\mathrm{i}\,\theta} - 1),$$

两端加上 $\frac{1}{2}(\mathrm{e}^{\mathrm{i}\,J\theta} - 1)$，得

$$\frac{1}{2} + \mathrm{e}^{\mathrm{i}\,\theta} + \mathrm{e}^{\mathrm{i}\,2\theta} + \cdots + \mathrm{e}^{\mathrm{i}\,(J-1)\theta} + \frac{1}{2}\mathrm{e}^{\mathrm{i}\,J\theta} = \frac{1}{2}(\mathrm{e}^{\mathrm{i}\,J\theta} - 1)\frac{(\mathrm{e}^{\mathrm{i}\,\theta} + 1)}{(\mathrm{e}^{\mathrm{i}\,\theta} - 1)} \tag{5.38}$$

$$= \frac{1}{2\mathrm{i}}(\mathrm{e}^{\mathrm{i}\,J\theta} - 1)\cot\frac{1}{2}\theta. \tag{5.39}$$

将此式与关于 $-\theta$ 的类似的和式相加即得 (5.37). 现将此式用于 $\theta=(k_1-k_2)\Delta x$, 因而有 $J\theta=(k_1-k_2)\pi$, 便得

$$\sum_{(j)}^{\prime} \mathrm{e}^{\mathrm{i}\,k_1 x_j}\mathrm{e}^{-\mathrm{i}\,k_2 x_j}=\sin(k_1-k_2)\pi\cot\frac{1}{2}(k_1-k_2)\Delta x,\quad k_1\neq k_2,$$

于是

$$\sum_{(j)}^{\prime} \mathrm{e}^{\mathrm{i}\,k_1 x_j}\mathrm{e}^{-\mathrm{i}\,k_2 x_j}=(2\pi/\Delta x)\delta_{k_1,k_2}.$$

■

因此，记 V_j 为控制体的测度，就有

$$\begin{aligned}\|U\|_2^2=\sum_{j\in J_\Omega}V_j|U_j|^2 &\equiv \sum_{(j)}^{\prime}(\Delta x)^d|U(\mathbf{x}_j)|^2\\ &=\left(\frac{\Delta x}{2\pi}\right)^d\sum_{(\mathbf{k})}^{\prime}|\widehat{U}(\mathbf{k})|^2\left(\frac{2\pi}{\Delta x}\right)^d,\end{aligned}\tag{5.40}$$

亦即

$$\|\widehat{U}\|_2^2:=\sum_{(\mathbf{k})}^{\prime}|\widehat{U}(\mathbf{k})|^2=\|U\|_2^2.\tag{5.41}$$

这就是 Parseval 关系式的恰当形式.

任意维空间上的矩形区域只需做一个简单的尺度变换即可化为以上情形. 不过，这时不仅 Δx 做了改变，且一般来说 $\Delta k\neq 1$, 要注意为了使 (5.41) 成立，在定义 $\|\widehat{U}\|_2$ 时需要用到这些系数. 同样值得注意的是，当考虑比如 [0, 1] 区间上具有边界条件 $u(0)=u(1)=0$ 的问题时，我们可以将问题在 $x=0$, 利用正弦函数序列反对称地延拓到 $[-1, 1]$ 区间. 这也是我们在上面取 $[-\pi, \pi]$ 为标准区间的原因.

为了对满足周期边界条件的常系数问题建立 (5.28) 式，先将任给的初始值按 (5.34) 的形式展开，然后，由 (5.28) 的离散傅里叶变换，得到相继各时间层上的同样形式的展开式，其系数为

$$\widehat{B}_1(\mathbf{k})\widehat{U}^{n+1}(\mathbf{k})=\widehat{B}_0(\mathbf{k})\widehat{U}^n(\mathbf{k}),\tag{5.42}$$

其中，当 U^n 为 p 维向量时， $\widehat{B}_0$ 和 $\widehat{B}_1$ 是 $p\times p$ 阶矩阵. 矩阵

$$G(\mathbf{k})=\widehat{B}_1^{-1}(\mathbf{k})\widehat{B}_0(\mathbf{k})\tag{5.43}$$

称为*增长矩阵*(amplification matrix), 因为它刻画了差分格式对每个波型的增长特性. 由于我们假设 $\widehat{B}_0$ 和 $\widehat{B}_1$ 不依赖时间 t, 所以有

$$\widehat{U}^n=[G(\mathbf{k})]^n\widehat{U}^0.\tag{5.44}$$

再结合 (5.41) 即得

$$\sup_{U^0}\frac{\|U^n\|_2}{\|U^0\|_2} = \sup_{\widehat{U}^0}\frac{[\sum'_{(\mathbf{k})}|\widehat{U}^n(\mathbf{k})|^2]^{1/2}}{[\sum'_{(\mathbf{k})}|\widehat{U}^0(\mathbf{k})|^2]^{1/2}}$$

$$= \sup_{\mathbf{k}}\sup_{\widehat{U}^0(\mathbf{k})}\frac{|\widehat{U}^n(\mathbf{k})|}{|\widehat{U}^0(\mathbf{k})|} = \sup_{\mathbf{k}}|[G(\mathbf{k})]^n|. \tag{5.45}$$

因此证明 l_2 范数意义下的稳定性就等价于证明

$$|[G(\mathbf{k})]^n| \leqslant K, \qquad \forall\mathbf{k},\ n\Delta t \leqslant t_F. \tag{5.46}$$

这里，$|G^n|$ 表示从属于 U_j^n 和 $\widehat{U}(\mathbf{k})$ 的向量范数的矩阵范数.

于是我们显然有以下结论.

定理 5.2. (冯诺伊曼定理) 稳定性的一个必要条件是存在常数 K' 使得对增长矩阵 $G(\mathbf{k})$ 的每个特征值 $\lambda(\mathbf{k})$ 都有

$$|\lambda(\mathbf{k})| \leqslant 1 + K'\Delta t, \qquad \forall\mathbf{k},\ n\Delta t \leqslant t_F. \tag{5.47}$$

证明. 将 $\widehat{U}(\mathbf{k})$ 取为 $G(\mathbf{k})$ 的某个特征向量，则显然一定存在常数 K 使得 $|\lambda^n| \leqslant K$; 取 $n\Delta t = t_F$ 即得 [1]

$$|\lambda| \leqslant K^{\Delta t/t_F} \leqslant 1 + (K-1)\Delta t/t_F, \qquad \Delta t \leqslant t_F,$$

其中最后一个不等式成立是由于 K^s 是 s 的凸函数. ∎

对标量方程的单步格式，若傅里叶分析可用于该格式，则 G 也是标量，这时冯诺伊曼条件是其 l_2 稳定性的充分条件. 当 G 是正规矩阵 (normal matrix) 时，冯诺伊曼条件也是其 l_2 稳定性的充分条件，因为此时 G 的谱半径是其从属范数的界. 在第 4 章中 Richtmyer 和 Morton (1967) 给出了其他更为复杂的充分条件，另外还有由 Kreiss[2] 和 Buchanan[3] 分别于 1962 年和 1963 年导出的充分必要条件. 主要的结果被称为 *Kreiss 矩阵定理*(Kreiss Matrix Theorem), 其内容已超出本书范围 (但可参见下面的 5.9 节). 不过值得指出的是，正如所见，当冯诺伊曼条件不能得到满足时，U^n 关于 n 可能会呈现指数增长，但当冯

[1] 原书下式有误：将 $\forall\Delta t \leqslant t_F/n$ 写成了 $\forall\Delta t \leqslant t_F$.

[2] Kreiss, H.O. (1962), Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren, *Nordisk Tidskr. Informations-Behandlung* **2**, 153-181.

[3] Buchanan, M.L. (1963), A necessary and sufficient condition for stability of difference schemes for second order initial value problems, *J. Soc. Indust. Appl. Math.* **11**, 474-501; 和 Buchanan, M.L. (1963), A necessary and sufficient condition for stability of difference schemes for initial value problems, *J. Soc. Indust. Appl. Math.* **11**, 919-35.

诺伊曼条件能得到满足，且 G 是一致有界时，U^n 的增长速度关于 n 最多是多项式的；事实上当 G 是 $p\times p$ 矩阵时，最坏的情况是 $O(n^{p-1})$ 的.

5.7 实用的(严厉的或强的)稳定性

显然，冯诺伊曼条件无论是在实际还是在理论上都十分重要，冯诺伊曼条件甚至于也可以局部地(取局部的系数值)应用到变系数问题. 由于高频波型是最不稳定的，因而不稳定性是一种局部现象，所以冯诺伊曼条件给出的必要条件也常常可以证明是充分的. 不过，对某些问题 (5.47) 中出现的任意常数在应用中显得过于宽泛，尽管对于最终的收敛性这已经足够了.

考虑以下一维扩散问题和对流问题的混合问题：

$$u_t + au_x = \epsilon u_{xx}, \qquad \epsilon > 0. \tag{5.48}$$

用空间中心差分和向前时间差分离散该问题：

$$\frac{U^{n+1}-U^n}{\Delta t} + a\frac{\Delta_{0x}U^n}{\Delta x} = \epsilon\frac{\delta_x^2 U^n}{(\Delta x)^2}. \tag{5.49}$$

现在我们有两个网格比

$$\nu := a\Delta t/\Delta x, \qquad \mu := \epsilon\Delta t/(\Delta x)^2, \tag{5.50}$$

傅里叶分析给出的增长因子为

$$\lambda(k) \;=\; 1 - \mathrm{i}\,\nu\sin k\Delta x - 4\mu\sin^2\frac{1}{2}k\Delta x, \tag{5.51a}$$

$$|\lambda|^2 \;=\; (1-4\mu s^2)^2 + 4\nu^2 s^2(1-s^2), \tag{5.51b}$$

其中采用了通常的记号 $s = \sin\frac{1}{2}k\Delta x$. 令 $s^2 = 1$, 得 $\mu\leqslant\frac{1}{2}$ 是稳定性的必要条件；这时

$$\nu^2 = (a\Delta t/\Delta x)^2 = (a^2/\epsilon)\mu\Delta t$$

而这意味着

$$|\lambda|^2 \leqslant 1 + \frac{1}{2}(a^2/\epsilon)\Delta t, \tag{5.52}$$

于是冯诺伊曼条件成立. 由于这是标量的纯初值问题，且 a 和 ϵ 都是常数，该条件也是稳定性的充分条件. 但若取 $\nu = 1$, $\mu = \frac{1}{4}$ 和 $s^2 = \frac{1}{2}$, 则有 $|\lambda|^2 = \frac{5}{4}$, 因此给出极快的增长；与此相比，原微分方程却会衰减掉所有的傅里叶波型.

在实际应用中，若冯诺伊曼条件所允许的指数增长与问题不再相适应，则对于 Δx 和 Δt 取有限值时，冯诺伊曼条件就显得太弱了. 因此我们需要引入以下更严格的定义.

定义 5.1. 一个格式称为是实用 (或严厉或强) 稳定的，如果当微分问题的傅里叶波型解对某个 $\alpha>0$ 满足

$$|\widehat{u}(\mathbf{k},t+\Delta t)|\leqslant \mathrm{e}^{\alpha\Delta t}|\widehat{u}(\mathbf{k},t)|,\qquad \forall\mathbf{k} \tag{5.53}$$

时，则对离散波型的所有 $\mathbf{k}$，差分格式相应的增长因子满足

$$|\lambda(\mathbf{k})|\leqslant \mathrm{e}^{\alpha\Delta t}. \tag{5.54}$$

对上面的例子，$\alpha=0$，因此要求 $|\lambda|\leqslant 1$. 由 (5.51b) 有

$$|\lambda|^2=1-4(2\mu-\nu^2)s^2+4(4\mu^2-\nu^2)s^4. \tag{5.55}$$

将此式看作为 s^2 在 [0, 1] 区间上的正的二次函数，即得到条件

$$|\lambda|^2\leqslant 1,\quad \forall k,\quad 当且仅当\quad \nu^2\leqslant 2\mu\leqslant 1. \tag{5.56}$$

这些条件是人们非常熟知的而且常常是重要的限制条件，因为当 ϵ 很小时这些条件会相当苛刻，除了预期中的条件 $\mu\leqslant\frac{1}{2}$，第一个不等式可以写为

$$\frac{a\cdot a\Delta t}{\epsilon}\equiv\frac{\nu^2}{\mu}\leqslant 2. \tag{5.57}$$

因此该条件可以解释为给网格 Péclet 数加以上限 2，在网格 Péclet 数的定义中 a 是速度，$a\Delta t$ 表示网格长度，而 ϵ 为扩散 (diffusion) 或粘性 (viscosity) 系数 (也见 2.15 节的讨论)，在 2.15 节中考虑最大值原理的应用时也出现过这类限制条件.

的确，这个稳定性的判据具有实际的重要性并导致对网格的限制，嵌入在 (5.53) 和 (5.54) 的稳定性定义是工程界经常使用的主要稳定性定义，但在使用该术语时并没有证明其在理论上的合理性. 注意在 5.9 节当应用于解不增长的问题时，该性质被称为*强稳定性* (strong stability)，相应于常微分方程离散化中绝对稳定性 (absolute stability) 的概念.

用以上同样的分析方法，将我们所分析过的任何一个扩散方程的显式格式与用于线性对流方程的格式相结合，可以得到类似实用的稳定性判据. 一般地说，所得到的条件比相应的两部分的条件组合在一起还要更加苛刻. 例如，将迎风格式 (4.13) 用于 (5.49) 代替中心差分，就得到

$$|\lambda|^2\leqslant 1,\quad \forall k,\quad 当且仅当\quad \nu^2\leqslant \nu+2\mu\leqslant 1, \tag{5.58}$$

而相对应的两个方程分别所需的条件为 $0\leqslant\nu\leqslant 1$ 和 $0\leqslant\mu\leqslant\frac{1}{2}$.

在做这些稳定性计算时，应用判别一个多项式的所有根是否全部落在闭单位圆盘上的一般判据会带来方便. Miller[1] 称这种多项式为冯诺伊曼多项式，并给出了利用那些相

[1] Miller, J. (1971), On the location of zeros of certain classes of polynomials with applications to numerical analysis, *J. Inst. Math. Appl.* **8**, 397-406.

关的 $m-1$ 次多项式来分析一个 m 次多项式是否为冯诺伊曼多项式的判据. 我们只给出复系数任意二次多项式的判据，该判据推广了我们在前几章中就一些特殊情形导出的结果.

引理 5.2. *系数为复数 a, b, c 的多项式 $a\lambda^2+2b\lambda+c=0$ 的所有根都满足条件 $|\lambda|\leqslant 1$, 当且仅当下列条件之一成立*

$$|c|<|a| \quad \text{且} \quad 2|\bar{a}b-\bar{b}c|\leqslant |a|^2-|c|^2, \tag{5.59a}$$

$$\text{或} \quad |c|=|a|,\ \bar{a}b=\bar{b}c \quad \text{且} \quad |b|\leqslant |a|. \tag{5.59b}$$

引理的证明留作习题 (见习题 3).

最后，我们再来回顾一下 2.11 节中给出的用于简单热流问题的 θ- 方法是稳定的或满足最大值原理所需的条件. 条件 $\mu(1-\theta)\leqslant \frac{1}{2}$ 给出了最大值原理，因而保证了最大模范数意义下的稳定性，且在最大模范数意义下误差无增长；另一方面，条件 $\mu(1-2\theta)\leqslant\frac{1}{2}$ 保证了 l_2 范数意义下的稳定性，在此之后我们还看到这也保证了取最大模范数时 (5.28) 意义下的稳定性. 然而，2.11 节中曾提到要使误差按最大模范数不增长的充分必要条件是 $\mu(1-\theta)^2\leqslant\frac{1}{4}(2-\theta)$, 注意，正如 (2.97) 中取 $K=1$ 时所表示的，这相应于最大模范数意义下的实用稳定性.

5.8 修正方程分析

这种类型的分析方法最初是作为推导稳定性条件的另外一种手段而引入的，不过在做此应用时必须非常小心. 现在人们认为修正方程分析 (modified equation analysis) 在获取有关数值方法的一般性态和逼近质量这类信息时更为有用. 这一术语是由 Warming 和 Hyett[1] 给出的. 不过，类似的思想更早由 Yanenko 和 Shokin[2] 以*微分逼近方法*(method of differential approximation) 为名提出，并由 Shokin (1983) 发展完善. 其基本思想是用网格值 $\{U_j^n\}$ 拟合出一个光滑函数，并找到该函数 (在达到用网格参数所定义的精度阶的意义下) 所满足的微分方程. 最近，类比于数值线性代数中通用的术语，这类方法又被称为*向后误差分析* (backward error analysis). 我们将通过几个例子来阐明这一思想.

要进行严格的分析，最好的办法也许是将网格数据做傅里叶展开，但较为简单的办法

[1] Warming, R.F. and Hyett, B.J. (1974), The modified equation approach to the stability and accuracy of finite difference methods, *J. of Comput. Phys.* **14**, 159-179.

[2] Yanenko, N.N. and Shokin, Y.I. (1969), First differential approximation method and approximate viscosity of difference schemes, *Phys. of Fluids* **12**, Suppl. II, 28-33.

是用多项式插值. 所需的分析运算与我们迄今为估计截断误差所做的十分相近. 因此, 先来考虑如 4.2 和 4.3 节中所述的用于线性对流方程的迎风格式. 记在网格点 (x_j, t_n) 的邻域中 U 的网格值的插值多项式为 $\tilde{U}(x,t)$. 将此代入差分方程 (4.13), 则类似于 (4.24) 可将两个差商展开为 (截断的) 泰勒级数

$$\left[\tilde{U}_t + \frac{1}{2}\Delta t\tilde{U}_{tt} + \cdots\right]_j^n + \left[a\left(\tilde{U}_x - \frac{1}{2}\Delta x\tilde{U}_{xx} + \cdots\right)\right]_j^n = 0. \tag{5.60}$$

根据 $\tilde{U}$ 所用的多项式的阶数, 将展开式在适当处截断, 即得到 $\tilde{U}(x,t)$ 所满足的**修正方程**(modified equation); 但是, 多项式的次数的选取也当然与其在相应网格点周围有效逼近的邻域密切相关. 作为一个极端的例子, 用格式中出现的三个值来拟合一个线性多项式, 则拟合出的线性多项式精确地满足原线性对流方程; 但这样做并没有为我们提供该差分格式逼近性态的任何整体信息.

因此, 一般将展开至更高阶, 具体阶数待定. 省略上下标后得到

$$\tilde{U}_t + a\tilde{U}_x = \frac{1}{2}\left[-\Delta t\tilde{U}_{tt} + a\Delta x\tilde{U}_{xx}\right] + \cdots. \tag{5.61}$$

这样的方程仍然用处不大, 因为右端项中包含关于时间的高阶导数. 解决这一问题的办法之一是在 (5.61) 上作用组合算子 $\partial_t - a\partial_x$, 若暂且假定 a 是正常数, 则由此给出的表达式将 $\tilde{U}_{tt}$ 用 $\tilde{U}_{xx}$ 表出. 将此表达式代入 (5.61) 的右端, 即得

$$\tilde{U}_t + a\tilde{U}_x = \frac{1}{2}\left[-a^2\Delta t + a\Delta x\right]\tilde{U}_{xx} + \cdots. \tag{5.62}$$

由此我们可以推出若取 $a\Delta t = \Delta x$, 则表明 $\tilde{U}$ 所满足的方程与作为目标的对流方程的差别至少是网格参量的二阶以上项; 又若取 $a\Delta t < \Delta x$, 则表明 $\tilde{U}$ 所满足的方程有一阶扰动, 且扰动是一个有阻尼效果的扩散项; 但若取 $a\Delta t > \Delta x$, 则扩散项的系数是负的, *这将导致严重的误差增长, 因而对应于不稳定性.*

比以上所用的方法更为有力的做法是有限差分演算 (例如参见 Hildebrand (1965) 第 5 章). 可以看出泰勒展开式中的系数恰恰是指数级数的系数, 因此可以将向前时间差分写成算子形式

$$\Delta_{+t} = e^{\Delta t\partial_t} - 1; \tag{5.63}$$

将此式形式求逆, 可以得到逆关系式

$$\partial_t = (1/\Delta t)\ln(1 + \Delta_{+t}). \tag{5.64}$$

然后由此式出发导出由 $\mathcal{D}_{+t} := \Delta_{+t}/\Delta t$ 表出的 ∂_t 的表达式,

$$\partial_t = \mathcal{D}_{+t} - \frac{1}{2}\Delta t\mathcal{D}_{+t}^2 + \frac{1}{3}(\Delta t)^2\mathcal{D}_{+t}^3 - \frac{1}{4}(\Delta t)^3\mathcal{D}_{+t}^4 + \cdots. \tag{5.65}$$

如果一个差分格式给出了用空间导数展开的 $\mathcal{D}_{+t}$ 的表达式, 则可立即得到所需的修正方程的展开式.

我们来考虑应用于对流扩散问题 (5.48) 的空间中心差分和时间向前差分的格式 (5.49). 展开空间差分算子，得

$$\mathcal{D}_{+t}\tilde{U} = \left\{-a\left[\partial_x + \frac{1}{6}(\Delta x)^2\partial_x^3 + \cdots\right] + \epsilon\left[\partial_x^2 + \frac{1}{12}(\Delta x)^2\partial_x^4 + \cdots\right]\right\}\tilde{U}. \tag{5.66}$$

然后，将此式代入展开式 (5.65), 合并同类项. 为了后面的应用，我们将结果展开至四阶空间导数项，得

$$\begin{aligned}\partial_t\tilde{U} \;=\; & \Big\{[(-a\partial_x + \epsilon\,\partial_x^2) + (\Delta x)^2(-\frac{1}{6}a\partial_x^3 + \frac{1}{12}\epsilon\,\partial_x^4 - \cdots)] \\ & -\frac{1}{2}\Delta t[a^2\partial_x^2 - 2a\epsilon\,\partial_x^3 + (\epsilon^2 + \frac{1}{3}a^2(\Delta x)^2)\partial_x^4 + \cdots] \\ & +\frac{1}{3}(\Delta t)^2[-a^3\partial_x^3 + 3a^2\epsilon\,\partial_x^4 + \cdots] \\ & -\frac{1}{4}(\Delta t)^3[a^4\partial_x^4 + \cdots] + \cdots\Big\}\tilde{U},\end{aligned} \tag{5.67}$$

此式可以写成修正方程

$$\tilde{U}_t + a\tilde{U}_x = \left[\epsilon - \frac{1}{2}a^2\Delta t\right]\tilde{U}_{xx} - \frac{1}{6}\left[a(\Delta x)^2 - 6a\,\epsilon\Delta t + 2a^3(\Delta t)^2\right]\tilde{U}_{xxx} + \cdots. \tag{5.68}$$

由此可以立即看出，不论如何选取时间步长，扩散系数都比原方程小，且仅当

$$\frac{1}{2}a^2\Delta t \leqslant \epsilon, \quad \text{或} \quad \frac{a\cdot(a\Delta t)}{\epsilon} \leqslant 2$$

时，扩散系数是非负的. 这是与长度标度为 $a\Delta t$ 的网格 Péclet 数小于 2 的要求相对应的稳定性的最大限度；这与由 (5.50) 式定义的网格比的术语给出的包含在 (5.56) 和 (5.57) 两式中的实用稳定性条件 $\nu^2 \leqslant 2\mu$ 是等价的. 注意，(5.68) 中的下一项可以用来估计格式的色散 (dispersion) 程度.

在这里我们应该指出，对于如上所得的修正方程的解释主要得益于傅里叶分析. 对于典型的傅里叶波型 $\mathrm{e}^{\mathrm{i}kx}$, 求 m 阶空间导数的结果就是乘以一个因子 $(\mathrm{i}k)^m$. 因此，奇数阶导数改变了波型的相位 (从而导致色散), 只有偶数阶导数改变了波型的大小并因此可能导致不稳定性. 如果二阶导数项的系数是负的 (前面刚讨论过的例子表明这种情形的确会出现) 则可推出不稳定性，因为当 $k\Delta x$ 很小时，该项就会控制住所有其他更高阶的项. 如果二阶导数项的系数为零，则需要进一步考虑更高阶的项，下面我们就来做这件事.

我们比较详细地给出了以上分析过程，因为这种分析可以用来比较任何一种显式的三点格式和对流扩散问题的形态. 为保证相容性，一阶差分项应为 $a\Delta_{0x}/\Delta x$, 但二阶差分项可以有不同的系数. 因此假设其系数由 ϵ 变到 ϵ'. 则由改变过的 (5.67) 式知，若令 $\epsilon' - \frac{1}{2}a^2\Delta t = \epsilon$, 则扩散项就能被正确模拟. 亦即 $\epsilon'\Delta t/(\Delta x)^2 = \epsilon\Delta t/(\Delta x)^2 + \frac{1}{2}(a\Delta t/\Delta x)^2$, 这

恰巧与 Lax-Wendroff 格式所给出的完全一样. 于是误差主项为色散项, $\tilde{U}_{xxx}$ 的系数为

$$-\frac{1}{6}a(\Delta x)^2 + a\Delta t\,(\epsilon + \frac{1}{2}a^2\Delta t) - \frac{1}{3}a^3(\Delta t)^2 = -\frac{1}{6}a(\Delta x)^2[1-(a\Delta t/\Delta x)^2] + a\,\epsilon\Delta t. \tag{5.69}$$

由此知, 对纯对流问题, Lax-Wendroff 格式的色散误差在任何稳定的网格比下总是负的, 上式在形式和量值上均与 4.5 节所示的该格式居主导地位的相位迟滞误差相符.

要分析 Lax-Wendroff 格式的稳定性, 需要考虑展开式 (5.67) 中包含四阶空间导数的项. 我们仅限于考虑 $\epsilon = 0$, 即纯对流问题. 因此, 记 CFL 数为 $\nu = a\Delta t/\Delta x$, 不难得出该方法的修正方程为

$$\tilde{U}_t + a\tilde{U}_x = -\frac{1}{6}\,a(\Delta x)^2[1-\nu^2]\tilde{U}_{xxx} - \frac{1}{8}\,a^2\Delta t(\Delta x)^2[1-\nu^2]\tilde{U}_{xxxx} + \cdots. \tag{5.70}$$

从而 $|\nu| > 1$ 时, 四阶导数项的系数变为正的, 对应于不稳定性.

这是一个很好的例子, 其中二阶项的系数为零, 但四阶导数项的系数可以变成正的, 于是对很小的 $k\Delta x$ 四阶项成为控制项, 我们因此可以推出不稳定性. 不过, 如果二阶项和四阶项的系数都是正的, 则情况就复杂得多: 我们会预期二阶项会起主导作用, 因此预示着稳定性; 但是我们必须考虑网格所能承载的所有傅里叶波型 (即满足 $|k\Delta x| \leqslant \pi$ 的所有波型) 以及修正方程的所有项.

基于以上对稳定性的考虑, 我们将注意力集中于纯对流问题. 因为不难看出即使对简单的热流问题修正方程方法也会失效. 例如, 假设在 (5.67) 式中令 $a = 0$, 我们要推导熟知的稳定性条件. 我们会发现情况正如上段所描述的那样, 很难推出一个合理的结论. 两类方程之间产生这种明显差别的基本原因在前面的傅里叶分析中已经看得很清楚. 许多对流方程的逼近方法的不稳定性表现在低频部分 (low frequencie), 类似于图 4-18 作增长因子图, 就会看到当 $k\Delta x$ 从零开始增大时, 增长因子就会移动到单位圆外; 对于这样的波型像 (5.67) 那样按空间导数阶的增序做展开是很有道理的. 但是, 正如 2.7 和 2.10 节所示, 热流问题典型的逼近方法的增长因子是实的, 且不稳定性首先出现在振荡最厉害的 (most oscillatory) 波型 $k\Delta x = \pi$ 处, 相应的增长因子小于 -1. 这时象 (5.67) 那样的展开式就毫无用处.

不过有办法破解这一两难的困境. 对每一个在 $|k\Delta x| \leqslant \pi$ 范围内的波型, 可将其写为 $k\Delta x = \pi - k'\Delta x$, 其中 $|k'\Delta x|$ 对大多数振荡最厉害的波型是小量; 于是关于该量做幂级数展开, 就相当于按振荡波型幅度的空间导数做展开. 下式等价于对傅里叶波型做分解

$$U_j^n = (U^s)_j^n + (-1)^{j+n}(U^o)_j^n, \tag{5.71}$$

亦即将 U 分解为光滑部分 U^s 和振荡部分 U^o, 并试图找到描绘后者性质的修正方程. 注意我们已经取出了关于时间变量的因子 $(-1)^n$, 因为那些变得不稳定的波型的增长因子预计会在 -1 附近.

现在来考虑与上面 (5.66) 相同的格式，不过是对 $a=0$ 给出的纯扩散问题. 对于振荡波型，提出公因子 $(-1)^{j+n}$, 得

$$\begin{aligned}-(U^o)_j^{n+1} &= (U^o)_j^n + \mu\left[-(U^o)_{j-1}^n - 2(U^o)_j^n - (U^o)_{j+1}^n\right] \\ &= (1-2\mu)(U^o)_j^n - 2\mu\left[1+\frac{1}{2}(\Delta x)^2\partial_x^2 + \frac{1}{24}(\Delta x)^4\partial_x^4 + \cdots\right](U^o)_j^n, \quad (5.72)\end{aligned}$$

其中方括号中的展开式是 $j=\pm1$ 时两项展开式的平均值. 因此，将所有这些项搜集到一起，就得到向前差商格式

$$\mathcal{D}_{+t}U^o = \left\{2(\Delta t)^{-1}(2\mu-1) + \epsilon\left[\partial_x^2 + \frac{1}{12}(\Delta x)^2\partial_x^4 + \cdots\right]\right\}U^o. \quad (5.73)$$

我们立即就可以看出，当 $\mu > \frac{1}{2}$ 时会有指数增长.

同样的方法可以帮助我们理解图 2.7 中的一些不稳定现象. 写出 Crank-Nicolson 格式的对等于 (5.73) 的表达式，则不难看出

$$\mathcal{D}_{+t}U^o = \frac{-2}{(2\mu+1)\Delta t}U^o + O((\Delta x)^2). \quad (5.74)$$

这显然是稳定的，因为对任意的正值 μ，它给出了按指数衰减的项. 当 Δt 趋向于零时，除了对极小的 t, 指数因子就小得微不足道了. 不过，在粗网格上情况就不一样了. 例如，当 $J=10$, $\mu=10$, 我们有 $\Delta t=1/10$, 所以指数因子为 $\exp(-(20/21)t)$, 这在整个 $0.1\leqslant t\leqslant 1$ 区间上都是相当大的. 这表明振荡项对该区间上按最大模范数的误差有着重大的影响，当网格加密时，这种影响迅速减小，以上分析表明，当 J 增加至 50 左右时，振荡项的影响就减少到可以忽略了.

在本节的最后，我们将修正方程分析应用于 4.8 节的盒式格式，其中 $(-1)^{j+n}$ 波型是著名的棋盘波型 (chequerboard mode), 它也是差分格式只由边界条件制约的伪解 (spurious solution). 较方便的做法是使用平均算子

$$\mu_x U_{j+1/2} := \frac{1}{2}(U_j + U_{j+1}), \quad (5.75)$$

类似地也有 μ_t, 利用这两个算子可以将线性 (常系数) 对流方程的盒式格式写为

$$\mu_x\delta_t U_{j+1/2}^{n+1/2} + (a\Delta t/\Delta x)\mu_t\delta_x U_{j+1/2}^{n+1/2} = 0. \quad (5.76)$$

我们可以将四个网格点上的值在它们的中点做泰勒展开，然后用类似于从 (5.61) 导出 (5.62) 的方法，将关于时间的高阶导数替换为空间的导数，最后推出下面得到的修正方程. 但下面要进行的是一种更具启发性且更为精巧的做法，我们先在盒式格式 (5.76) 上作用两个平均算子的逆算子，并引入差分算子

$$\mathcal{D}_x = \frac{\mu_x^{-1}\delta_x}{\Delta x}, \qquad \mathcal{D}_t = \frac{\mu_t^{-1}\delta_t}{\Delta t}. \quad (5.77)$$

这样我们就可以将该格式写成紧凑的形式 $(\mathcal{D}_t + a\mathcal{D}_x)U = 0$. 尽管这一步只是形式上的运算，实际计算时并不会这么做，但它却显示出盒式格式是用一种相容的方式将微分算子替换为差分算子，而且由此导出了与微分方程直接对应的差分关系式. 更重要的是，这给出了一种利用算子演算 (operator calculus) 推导修正方程的直接且极具普遍性的方法.

第一步是按类似于 (5.63) 和 (5.64) 的做法得到差分算子关系式

$$\mathcal{D}_x = (\frac{1}{2}\Delta x)^{-1}\tanh(\frac{1}{2}\Delta x\partial_x),$$

由此可推出

$$\mathcal{D}_x = \left[1 - \frac{1}{12}(\Delta x)^2\partial_x^2 + \cdots\right]\partial_x. \tag{5.78}$$

对时间差分算子做类似的演算，然后取逆，得

$$\partial_t = \left[1 + \frac{1}{12}(\Delta t)^2\mathcal{D}_t^2 + \cdots\right]\mathcal{D}_t. \tag{5.79}$$

应用 (5.78, 5.79) 两个展开式，就立即得到了格式的修正方程

$$\tilde{U}_t + a\tilde{U}_x = \frac{1}{12}a(\Delta x)^2(1-\nu^2)\tilde{U}_{xxx} + \cdots, \tag{5.80}$$

这里我们引入了 CFL 数 ν.

展开式中的所有其他项都只含奇数阶导数，这与格式的无条件稳定性相一致；色散项的系数根据 $\nu < 1$ (导致相位超前 (phase advance)) 或 $\nu > 1$ (导致相位滞后 (retardation)) 而改变符号. 事实上，展开式 (5.80) 与 (4.85) 所给出的增长因子的相位展开式完全精确地相对应. 不过修正方程可以推广到传播速度为变量的问题，如图 4-13 所给出的例子，事实上还可以推广到非线性问题.

修正方程 (5.80) 仅适用于流的光滑分量. 但图 4-13 显示间断的初始值会引起解的剧烈振荡. 因此，我们现在来推导对应于分解式 (5.71) 中振荡分量的修正方程. 对正在使用的差分算子，不难看出对分解式 (5.71) 中振荡分量有

$$\begin{aligned}\delta_x U_{j+1/2} &= (-1)^{j+1}2\mu_x U^o_{j+1/2},\\ \mu_x U_{j+1/2} &= (-1)^{j+1}\frac{1}{2}\delta_x U^o_{j+1/2},\end{aligned} \tag{5.81}$$

即差分变成平均，而平均变成差分. 因此，由 (5.76), U^o 满足的方程为

$$\mu_t\delta_x(U^o)^{n+1/2}_{j+1/2} + (a\Delta t/\Delta x)\mu_x\delta_t(U^o)^{n+1/2}_{j+1/2} = 0. \tag{5.82}$$

比较两个方程就可以立即写出振荡分量所满足的修正方程,容易看出,若令 $\nu_2 = a(\Delta t/\Delta x)^2$ 并考虑将 a 换为 $1/\nu_2$ 后的对流扩散方程，则 U^o 满足关于该方程的盒式格式，于是有

$$\tilde{U}^o_t + (1/\nu_2)\tilde{U}^o_x = \frac{1}{12}(1/\nu_2)(\Delta x)^2(1-\nu^{-2})\tilde{U}^o_{xxx} + \cdots, \tag{5.83}$$

因此这些波型依然没有被衰减，而是以完全不同于 a 的依赖于网格比的速度传播．如 4.8 节所示，这个传播速度也可以由这些波型的群速度推出， $C_h = a/\nu^2 \equiv 1/\nu_2$.

4.8 节还提到，棋盘波型通常可用加权时间平均的方法控制．所以我们最后来看一下这种做法对修正方程的效果．定义平均算子

$$\theta_t U^{n+1/2} := \theta U^{n+1} + (1-\theta)U^n, \tag{5.84}$$

因而有 $\theta_t = \mu_t + (\theta - \frac{1}{2})\delta_t$. 然后引入算子

$$\mathcal{M}_t := \mu_t^{-1}\theta_t = 1 + (\theta - \frac{1}{2})\Delta t\, \mathcal{D}_t. \tag{5.85}$$

利用该算子，可将加权盒式格式改写成

$$[\mathcal{D}_t + a\mathcal{M}_t\mathcal{D}_x]U_{j+1/2}^{n+1/2} = 0. \tag{5.86}$$

记 $\gamma = (\theta - \frac{1}{2})\Delta t$, 求解出时间方向的差分，便得到展开式

$$\begin{aligned} \mathcal{D}_t U &= (1+\gamma a\,\mathcal{D}_x)^{-1}(-a\,\mathcal{D}_x)U \\ &= \left[-a\,\mathcal{D}_x + \gamma a^2\mathcal{D}_x^2 - \gamma^2 a^3\mathcal{D}_x^3 + \cdots\right] U. \end{aligned} \tag{5.87}$$

然后代入 (5.78, 5.79) ，就得修正方程

$$\tilde{U}_t + a\tilde{U}_x = \gamma a^2 \tilde{U}_{xx} + a(\Delta x)^2 \left[\frac{1}{12}(1-\nu^2) - \gamma^2 a^2\right] \tilde{U}_{xxx} + \cdots . \tag{5.88}$$

因此，若取 $\theta < \frac{1}{2}$, 则有 $\gamma < 0$ ，从而导致不稳定性；但若取 $\theta > \frac{1}{2}$ ，则引入了阻尼．通常 θ 的取值限制在 $\theta = \frac{1}{2} + O(\Delta t)$.

再来考虑振荡波型．由 (5.81), 我们将 (5.86) 变为

$$\left\{\mathcal{D}_x + a(\Delta t/\Delta x)^2 \left[\mathcal{D}_t + 4(\Delta t)^{-1}(\theta - \frac{1}{2})\right]\right\} U^o = 0, \tag{5.89}$$

由此，便得到修正方程

$$\tilde{U}_t^o + (1/\nu_2)\tilde{U}_x^o = -4(\Delta t)^{-1}(\theta - \frac{1}{2})\tilde{U}^o + \cdots . \tag{5.90}$$

因此，甚至取 $\theta = \frac{1}{2} + O(\Delta t)$ 时，只要 $\theta > \frac{1}{2}$, 则棋盘波型也是指数衰减的．

5.9 守恒律与能量法分析

我们已经在 2.7 和 4.11 节中看到，关于 $\int u\,\mathrm{d}x$ 的守恒律可以帮助选取边界条件．现在我们来考虑如何利用类似的关于 $\int |u|^2 \mathrm{d}x$ 的守恒律，推导出建立稳定性的有用方法．这类方法的起源及思想有若干出处．无可质疑最为重要的是在建立写成为 (5.1a) 形式的一大类偏微分方程的适定性时，能量不等式所起的作用．设将方程的两端分别与 u 作内

积，在 Ω 上积分并利用格林公式 (Green's Identities), 则在一定的边界条件下有

$$(u,\ Lu)_2 \equiv \int_\Omega uL(u)\,\mathrm{d}\Omega \leqslant K\int_\Omega |u|^2\mathrm{d}\Omega \equiv K\|u\|_2^2. \tag{5.91}$$

由此就可以建立微分问题在 L_2 范数意义下的适定性. 因为该范数有时表示的是系统的物理能量，所以不论是应用于微分问题还是其有限差分近似，这类方法一般被称为*能量方法* (energy method).

作为一个简单的例子，我们先回到第二章最开始所引入的模型问题，即带齐次狄利克雷边界条件的一维热传导方程：

$$u_t = u_{xx}, \qquad 0<x<1,\ \ t>0, \tag{5.92a}$$

$$u(0,t) = u(1,t)=0, \qquad t>0, \tag{5.92b}$$

$$u(x,0) = u^0(x), \qquad 0<x<1. \tag{5.92c}$$

由 (5.92a), 立即得到 $(\frac{1}{2}u^2)_t = u\,u_{xx}$, 因此做分部积分后得

$$\frac{\partial}{\partial t}\left(\int_0^1 \frac{1}{2}u^2\mathrm{d}x\right) = \int_0^1 u\,u_{xx}\mathrm{d}x = [u\,u_x]_0^1 - \int_0^1 (u_x)^2\mathrm{d}x \tag{5.93}$$

$$= -\int_0^1 (u_x)^2\mathrm{d}x \leqslant 0, \tag{5.94}$$

其中我们还用到了边界条件 (5.92b). 因此我们证明了 u 的 L_2 范数是递减的，因而被其初始值所界定. 注意，在这种情况下， u 的 L_2 范数并不对应于系统的物理能量.

在这个分析中最主要的工具是分部积分. 要得到离散情形的类似结果需要用到下面将要给出的*分部求和* (summation by parts) 公式. 但是我们首先指出离散情形的这种分析方法，直接受到曾在 5.6 节提到过的 Kreiss 矩阵定理的启发. 该定理的其中一个命题，可以证明与 (5.46) 式所表述的增长矩阵 G 的幂有界性等价，该命题称存在能量范数 $|v|_H^2 := v^*Hv$, 使得 $|G|_H \leqslant 1$; 即存在一致有界正定的埃尔米特矩阵 (Hermitian matrix)H 满足

$$G^*HG \leqslant H. \tag{5.95}$$

因此在该范数意义下，每一时间步误差都不会增长. 不幸地是我们一般并不知道如何构造或选取这种矩阵 H, 或者当我们从傅里叶空间转换到实空间时，不知道如何构造或选取相应的算子. 而我们实际需要的正是这些实空间的算子，因为该定理最重要的目标就是要把由傅里叶分析得到的稳定性结果推广到变系数、非周期边界条件等情形. 另一方面，对一些重要的问题，我们能够凭感觉通过组合一些差分算子构造出这种算子. 本节中我们将通过若干例子来演示如何做到这一点. 得到这种算子之后，就可以建立所考虑的差分格式的稳定性了，但是我们却无法证明不稳定性. 因此这种方法给出的是稳定性的充分条件，这些条件补充了由冯诺伊曼分析给出的必要条件.

在开始做这件事之前，先来介绍一个最近的结果[1]，该结果使我们能够非常直接地证明一大类重要方法的稳定性. 该定理事实上给出了与我们定义的实用稳定性等价的 强稳定性(strong stability) 的条件，因为它能够直接适用的仅限于解不会增长的问题.

定理 5.3. 设通过（反复）应用一个空间差分算子 L_Δ，并将其写成称之为 **Runge-Kutta** 时间步进法的展开式的形式，来近似计算一个 (5.1a) 形式的适定问题：

$$U^{n+1}=\sum_{i=0}^{s}\frac{(\Delta tL_\Delta)^i}{i!}U^n, \tag{5.96}$$

又设算子 L_Δ 在下述意义下是强制的 **(coercive)**，即存在正常数 η，使得所有的网格函数 U 都满足

$$(L_\Delta U,\ U)\leqslant -\eta\|L_\Delta U\|^2. \tag{5.97}$$

则对 $s=1,2,3,$ 或 4, 当时间步长满足条件

$$\Delta t\leqslant 2\eta \tag{5.98}$$

时，格式是强稳定的.

证明. 对一个不反复应用该算子的方法，即 $s=1$, 且算子在 (5.97) 的意义下是强制的，有

$$\begin{aligned}\|U^{n+1}\|^2 &= \|U^n+\Delta t\,L_\Delta U^n\|^2\\ &= \|U^n\|^2+2\Delta t\,(U^n,\ L_\Delta U^n)+(\Delta t)^2\|L_\Delta U^n\|^2\\ &\leqslant \|U^n\|^2+\Delta t\,(\Delta t-2\eta)\|L_\Delta U^n\|^2.\end{aligned} \tag{5.99}$$

因此，当 (5.98) 满足时，近似解的范数不增加. 对 $s=2$，不难验证

$$I+\Delta tL_\Delta+\frac{1}{2}(\Delta tL_\Delta)^2=\frac{1}{2}I+\frac{1}{2}(I+\Delta tL_\Delta)^2,$$

对 $s=3$, 可类似地验证

$$\sum_{i=0}^{3}\frac{(\Delta tL_\Delta)^i}{i!}=\frac{1}{3}I+\frac{1}{2}(I+\Delta tL_\Delta)+\frac{1}{6}(I+\Delta tL_\Delta)^3,$$

而对 $s=4$, 有

$$\sum_{i=0}^{4}\frac{(\Delta tL_\Delta)^i}{i!}=\frac{3}{8}I+\frac{1}{3}(I+\Delta tL_\Delta)+\frac{1}{4}(I+\Delta tL_\Delta)^2+\frac{1}{24}(I+\Delta tL_\Delta)^4.$$

由于我们已经证明了当 (5.98) 满足时，有 $\|I+\Delta tL_\Delta\|\leqslant 1$, 因此该算子的所有幂次都满足类似的不等式；又因为以上每个展开式的各项系数之和为 1, 于是 1 是所有这三个算子展开式共同的界. ■

[1] Levy, D. and Tadmor, E. (1998), From semi-discrete to fully discrete: stability of Runge-Kutta schemes by the energy method, *SIAM Rev.* **40**, 40-73; and Gottlieb, S., Shu, C.-W. and Tadmor, E. (2001), Strong stability-preserving high-oder time discretization methods, *SIAM Rev.* **43**, 89-112.

在应用上述定理之前，我们先来作一些评注．定理的思想来源于**半离散方法** (semi-discrete method); 即首先将空间做离散化，把偏微分方程转化为大规模的常微分方程组 (large system of ODEs), 然后应用例如 Runge-Kutta 时间步进法求解．对于 $s>1$ 的情形，显然必须注意使用适当的边界条件；要将这种思想用于非线性问题，则有更多需要注意的地方．但是，我们在这里仅局限于讨论标准的线性问题，且仅将定理用于 $s=1$ 的情形．不过，这仍然是一个应用能量分析建立若干方法的稳定性的有用的框架．

定理中的实内积和范数可以相当一般，而我们的目的是将定理的结果用于离散的 l_2 范数．其一般形式为 (5.5), 但我们将例子限制在一维，并且假设采用的是一致网格，记 $J_\Omega=\{1,2,\cdots,J-1\}$, 这时两个实向量 U 和 V 的 l_2 内积为

$$\langle U,\ V\rangle_2=\Delta x\sum_1^{J-1}U_jV_j, \tag{5.100}$$

相应的范数为 $\|U\|_2=\langle U,\ U\rangle_2^{1/2}$. 在相应的网格上分部求和公式由以下引理给出．

引理 5.3. 对任意数列 $\{V_j\}$ 和 $\{W_j\}$, 有

$$\sum_1^{J-1}[V_j(W_{j+1}-W_j)+W_j(V_j-V_{j-1})]=V_{J-1}W_J-V_0W_1 \tag{5.101}$$

和

$$\sum_1^{J-1}[V_j(W_{j+1}-W_{j-1})+W_j(V_{j+1}-V_{j-1})]=(V_{J-1}W_J+W_{J-1}V_J)-(V_0W_1+W_0V_1). \tag{5.102}$$

直接进行代数运算即可证明 (5.101), 具体过程留作习题．在 (5.101) 中交换 V 和 W 的位置，再将所得的两个结果相加即得第二个公式 (5.102).

用差分算子可以将 (5.101) 和 (5.102) 写成

$$\langle V,\ \Delta_{+x}W\rangle_2+\langle W,\ \Delta_{-x}V\rangle_2=\Delta x(V_{J-1}W_J-V_0W_1) \tag{5.103}$$

和

$$\langle V,\ \Delta_{0x}W\rangle_2+\langle W,\ \Delta_{0x}V\rangle_2=\frac{1}{2}\Delta x\left[(V_{J-1}W_J+W_{J-1}V_J)-(V_0W_1+W_0V_1)\right]. \tag{5.104}$$

在第一个式子中用 $\Delta_{-x}W$ 替换 W , 还可以得到

$$\langle V,\ \delta_x^2W\rangle_2+\langle\Delta_{-x}W,\ \Delta_{-x}V\rangle_2=\Delta x\left[V_{J-1}(W_J-W_{J-1})-V_0(W_1-W_0)\right]; \tag{5.105}$$

因此，如果有 $V=W$ 及 $V_0=V_J=0$, 则有

$$\langle V,\ \delta_x^2V\rangle_2=-\|\delta_xV\|_2^2:=-\Delta x\sum_{j=1}^{J}|\delta_xV_{j-1/2}|^2, \tag{5.106}$$

其中应该注意的是，求和范围包括了所有 J 个单元．

例 1. 首先考虑应用显式格式求解 (5.92). 格式可以写成

$$U_j^{n+1} = U_j^n + \mu\delta_x^2 U_j^n, \tag{5.107}$$

其中 $\mu = \Delta t/(\Delta x)^2$. 现将定理 5.3 应用于该格式，这里 $L_\Delta = (\Delta x)^{-2}\delta_x^2$. 由 (5.106) 有

$$(\Delta x)^2(L_\Delta U,\ U) = \langle\delta_x^2 U,\ U\rangle = -\|\delta_x U\|^2, \tag{5.108}$$

在此式中，我们省略了范数 $\|\cdot\|_2$ 和内积 $\langle\cdot,\ \cdot\rangle_2$ 的下标 (本节剩下的部分里也将做同样的省略); 由 Cauchy-Schwarz 不等式，我们还有

$$\begin{aligned}(\Delta x)^4\|L_\Delta U\|^2 &= \|\delta_x^2 U\|^2 = \Delta x\sum_{j=1}^{J-1}(\delta_x U_{j+1/2} - \delta_x U_{j-1/2})^2 \\ &\leqslant 2\Delta x\sum_{j=1}^{J-1}\left[(\delta_x U_{j+1/2})^2 + (\delta_x U_{j-1/2})^2\right] \\ &\leqslant 4\,\|\delta_x U\|^2.\end{aligned} \tag{5.109}$$

比较 (5.108) 和 (5.109), 显然定理 5.3 对 $\eta = \frac{1}{4}(\Delta x)^2$ 成立，于是推出当 $\Delta t \leqslant \frac{1}{2}(\Delta x)^2$ 时，格式具有稳定性的熟知的结果.

例 2. 其次我们考虑应用隐式 θ- 格式求解同样的问题. 定理 5.3 不适用于该格式. 因此我们推导如下，将差分方程写为

$$U_j^n - U_j^{n-1} = \mu\left[\theta\delta_x^2 U_j^n + (1-\theta)\delta_x^2 U_j^{n-1}\right], \tag{5.110}$$

在上式两端同乘以 $U_j^n + U_j^{n-1}$, 并对使上式成立的所有 j , 即 $j = 1, 2, \ldots, J-1$, 求和. 于是得到

$$\|U^n\|^2 - \|U^{n-1}\|^2 = \mu\langle U^n + U^{n-1},\ \delta_x^2[\theta U^n + (1-\theta)U^{n-1}]\rangle.$$

对上式的右端应用分部求和公式 (5.105), 然后再应用形如 $-ab \leqslant \frac{1}{2}(a^2+b^2)$ 的 Cauchy-Schwarz 不等式，就得到

$$\begin{aligned}\|U^n\|^2 - \|U^{n-1}\|^2 &= -\mu\langle\delta_x[U^n + U^{n-1}],\ \delta_x[\theta U^n + (1-\theta)U^{n-1}]\rangle \\ &= -\mu\left\{\theta\|\delta_x U^n\|^2 + (1-\theta)\|\delta_x U^{n-1}\|^2 + \langle\delta_x U^n,\ \delta_x U^{n-1}\rangle\right\} \\ &\leqslant -\mu\left\{(\theta - \frac{1}{2})\|\delta_x U^n\|^2 + (\frac{1}{2} - \theta)\|\delta_x U^{n-1}\|^2\right\}.\end{aligned} \tag{5.111}$$

注意最后一个表达式中两项的系数大小相等符号相反. 因此，当将同样的公式用于前面各步，并将结果求和时，则所有的中间项都会相互抵消. 易见，其结果是

$$\|U^n\|^2 - \|U^0\|^2 \leqslant -\mu(\theta - \frac{1}{2})\left\{\|\delta_x U^n\|^2 - \|\delta_x U^0\|^2\right\}.$$

我们将此式改写为更有意义的表达式

$$\|U^n\|^2 + \mu(\theta - \frac{1}{2})\|\delta_x U^n\|^2 \leqslant \|U^0\|^2 + \mu(\theta - \frac{1}{2})\|\delta_x U^0\|^2. \tag{5.112}$$

如果上式两端的表达式对应于某种范数的话，则稳定性也就隐含其中了. 事实上，若上式两端的第一项都能控制住第二项，则格式具有 l_2 范数 [1] 意义下的稳定性. 在 (5.109) 中，Cauchy-Schwarz 不等式被用来导出 $\delta_x^2 U$ 的范数的界，现在需要同样地导出 $\delta_x U$ 的范数的界；实际上，用 (5.109) 中的方法可以得到有用的不等式

$$\|\delta_x U\|^2 \leqslant 4\,\|U\|^2. \tag{5.113}$$

将此用于 (5.112), 即推出 $\mu(1-2\theta) \leqslant \frac{1}{2}$, 给出稳定性，这与 2.10 节得到的结果几乎是等价的，这里的方法只是在该关系式为等式时没能推出稳定性，但这只对 $\theta < \frac{1}{2}$ 的情形有影响. 还要提请大家注意，因为由 (5.111) 知，当稳定性条件满足时该范数在每一步是非增的，所以这个例子说明，我们能够显式地构造出 Kreiss 矩阵定理所确保存在的范数.

例 3. 在本例及后面一个例子中我们将说明如何应用能量分析来处理更加复杂的问题，尤其是变系数的问题和需要施加数值边界条件的问题. 考虑以下问题

$$u_t + a(x)u_x = 0, \qquad 0 < x \leqslant 1,\ t > 0, \tag{5.114a}$$

$$u(x,0) = u^0(x), \qquad 0 \leqslant x \leqslant 1, \tag{5.114b}$$

$$u(0,t) = 0, \qquad t > 0. \tag{5.114c}$$

设 $a(x)$ 非负、有界且满足利普希茨 (Lipschitz) 条件：

$$0 \leqslant a(x) \leqslant A, \qquad |a(x) - a(y)| \leqslant K_L|x-y|. \tag{5.115}$$

注意，与特征线从左指向右相容，齐次狄利克雷边界条件加在了左端.

本例中，我们考虑使用一阶迎风格式，并采用非常直接的类似于证明定理 5.3 的方法. 令 $L_\Delta = -a(\Delta x)^{-1}\Delta_{-x}$, 则有

$$\begin{aligned}\langle L_\Delta U,\, U\rangle &= -\sum_{j=1}^{J-1} a_j(U_j - U_{j-1})U_j \\ &= -\sum_{j=1}^{J-1} a_j U_j^2 + \sum_{j=1}^{J-1} a_j U_j U_{j-1};\end{aligned} \tag{5.116}$$

和

$$\begin{aligned}\Delta x\,\|L_\Delta U\|^2 &= \sum_{j=1}^{J-1} a_j^2 (U_j - U_{j-1})^2 \leqslant A\sum_{j=1}^{J-1} a_j (U_j - U_{j-1})^2 \\ &= A\left\{\sum_{j=1}^{J-1} a_j U_j^2 - 2\sum_{j=1}^{J-1} a_j U_j U_{j-1} + \sum_{j=1}^{J-1} a_j U_{j-1}^2\right\}.\end{aligned} \tag{5.117}$$

[1] 原书中是 L_2 范数. 但显然这里用的是离散 l_2 范数.

于是推出

$$
\begin{aligned}
2\langle L_\Delta U,\, U\rangle + A^{-1}\Delta x\,\|L_\Delta U\|^2 &\leqslant -\sum_{j=1}^{J-1} a_j(U_j^2 - U_{j-1}^2) \\
&= -\sum_{j=1}^{J-2}(a_j - a_{j+1})U_j^2 - a_{J-1}U_{J-1}^2 \\
&\leqslant K_L\|U\|^2, \qquad (5.118)
\end{aligned}
$$

其中用到了左端边界上 U 的边界条件和 a 的利普希茨条件. 因此, 若稳定性条件 $A\Delta t \leqslant \Delta x$ 得以满足, 就有

$$
\begin{aligned}
2\Delta t\,\langle L_\Delta U,\, U\rangle + \Delta t^2\|L_\Delta U\|^2 &\leqslant \Delta t\,\left[2\langle L_\Delta U,\, U\rangle + A^{-1}\Delta x\,\|L_\Delta U\|^2\right] \\
&\leqslant K_L\Delta t\,\|U\|^2. \qquad (5.119)
\end{aligned}
$$

这就给出了结论

$$\|(I + \Delta t\, L_\Delta)U\|^2 \leqslant (1 + K_L\Delta t)\|U\|^2, \qquad (5.120)$$

亦即格式的稳定性. 本例利用能量方法证明变系数不会破坏稳定性, 尽管会引起误差增大, 这是具有代表性的方法.

例 4. 仍然考虑同样的对流问题 (5.114), 但应用蛙跳格式, 且为简单起见, 设 a 为正常数. 格式可以表示为

$$U_j^{n+1} - U_j^{n-1} = -2\nu\Delta_{0x}U_j^n, \qquad j \in J_\Omega \equiv \{1, 2, \ldots, J-1\}, \qquad (5.121)$$

其中 $\nu = a\Delta t/\Delta x$. 该格式给出了 U_j^{n+1} 的显式计算公式, 但仅限于内点. U_0^{n+1} 的值由边界条件给出, 而 U_J^{n+1} 的值则必须用其他方法计算. 我们将从格式稳定性的能量分析中推出适当的数值边界条件.

在 (5.121) 两端同乘以 $U_j^{n+1} + U_j^{n-1}$, 然后求和, 得

$$\|U^{n+1}\|^2 - \|U^{n-1}\|^2 = -2\nu\langle U^{n+1} + U^{n-1},\, \Delta_{0x}U^n\rangle. \qquad (5.122)$$

与例 2 相同, 我们要将 (5.122) 右端表示成为两个类似的表达式, 在相邻时间步取值之差的形式. 为此, 应用第二个分部求和公式 (5.104), 并在其中令 $V = U^{n-1}$ 及 $W = U^n$, 得

$$2\nu\left[\langle U^{n-1},\, \Delta_{0x}U^n\rangle + \langle U^n,\, \Delta_{0x}U^{n-1}\rangle\right] = a\Delta t\left[U_{J-1}^{n-1}U_J^n + U_J^{n-1}U_{J-1}^n\right]. \qquad (5.123)$$

结果就得到

$$
\begin{aligned}
\|U^{n+1}\|^2 - \|U^{n-1}\|^2 = {} & -2\nu\langle U^{n+1},\, \Delta_{0x}U^n\rangle + 2\nu\langle U^n,\, \Delta_{0x}U^{n-1}\rangle \\
& - a\Delta t\left[U_{J-1}^{n-1}U_J^n + U_J^{n-1}U_{J-1}^n\right]. \qquad (5.124)
\end{aligned}
$$

含内积的两项在关于时间逐步求和时会相互抵消；不过这两项中还包含右端边界点的值 U_J^n, 而我们还没有定义怎样计算该值. 将 (5.124) 中的所有边界项合并在一起，得

$$-a\Delta t\left[U_{J-1}^{n+1}U_J^n - U_{J-1}^n U_J^{n-1} + U_{J-1}^{n-1}U_J^n + U_J^{n-1}U_{J-1}^n\right].$$

其中第二和第四项相互抵消；因此，如果施加边界条件

$$U_J^n = \frac{1}{2}\left(U_{J-1}^{n-1} + U_{J-1}^{n+1}\right), \tag{5.125}$$

则剩余的边界项的贡献是负定的. 所以，如果将所得到的不等式在时间层 $n, n-1, \ldots, 2, 1$ 上求和，并在内积上加一撇表示从中去掉了右端边界项，就得到

$$\|U^{n+1}\|^2 + \|U^n\|^2 + 2\nu\langle U^{n+1}, \Delta_{0x}U^n\rangle' \leqslant \|U^1\|^2 + \|U^0\|^2 + 2\nu\langle U^1, \Delta_{0x}U^0\rangle'. \tag{5.126}$$

最后，对内积项应用 Cauchy-Schwarz 不等式，得

$$\begin{aligned} 2\langle U^{n+1}, \Delta_{0x}U^n\rangle' &= \Delta x\sum_{j=1}^{J-1} U_j^{n+1}[U_{j+1}^n - U_{j-1}^n]' \\ &\leqslant \|U^{n+1}\|^2 + \|U^n\|^2, \end{aligned} \tag{5.127}$$

其中，再次用一撇表示从求和中去掉了右端边界项. 由此知，若取 $\nu < 1$, 则 (5.126) 两端均为正定的，且与各自的 l_2 范数等价. 亦即对这样的 CFL 数，我们证明了 l_2 范数意义下的稳定性，但与例 2 一样我们不能用这种方法证明 $\nu = 1$ 时的稳定性.

5.10　理论综述

毫无疑问，傅里叶分析是研究差分格式稳定性以及精度的最有用和最精确的工具；而由 Lax 等价定理，稳定性则是最关键的性质. 然而，由于傅里叶分析仅能够应用于具有常系数且满足周期边界条件的线性问题，且只能研究定义在一致网格上的差分格式在 l_2 范数意义下的稳定性，因此傅里叶分析方法也是一种最受限制的方法. 所以，过去四十年有关这方面的大多数的理论进展，可以认为是在表明由傅里叶分析得到的结论可以有更广泛的应用. 认为拓广傅里叶分析应用范围是可行的想法源于早期冯诺伊曼的观察，即对于差分格式来说最不稳定的是高频波型，因此不稳定性更应该是一种局部的现象.

对抛物型问题，我们已经通过若干例子显示最大值原理可以用来证明很大一类问题 (包括变系数、非一致网格、非矩形区域和实际的边界条件等) 的稳定性. 傅里叶分析在其适用时给出的是稳定性的必要条件，这些条件通常比用最大模范数分析得到的条件弱. 不过，由前面曾经提到过的 Fritz John 的工作，傅里叶分析得到的条件在证明 (5.28) 意义下的稳定性时也是充分条件，而且由定理 5.1 知，为了保证收敛性需要用到这种稳定性.

由于两位作者的导致重大发展的综述论文[1]，这种稳定性也常常被称为 *Lax-Richtmyer 稳定性*(Lax-Richtmyer stability). 情形之所以是这样，其主要原因有两个.

第一个原因是 (5.28) 中的稳定性条件的选择本身就是为了得到这样的结果. 20 世纪 50 年代使用着许多不同的稳定性定义，这些稳定性定义会引发十分混乱的现象. 例如，一方面可以举出一些例子说明由局部导出的稳定性条件不是全局稳定性的充分条件，而另一方面又可以举出另一些例子说明这些条件不是必要的. 而 (5.28) 的关键性质是差分格式的小扰动不改变其稳定性. 基于在 5.6 节曾提及的 Kreiss (1962) 和 Strang (1964)[2] 关于这些稳定性现象的研究，我们有以下引理.

引理 5.4. *设差分格式 $U^{n+1}=B_1^{-1}B_0U^n$ 是稳定的，且 $C(\Delta t)$ 是一族有界算子. 则格式*

$$U^{n+1}=\left[B_1^{-1}B_0+\Delta t\,C(\Delta t)\right]U^n \tag{5.128}$$

是稳定的.

证明. 设 $\|(B_1^{-1}B_0)^n\|\leqslant K_1$, $\|C(\Delta t)\|\leqslant K_2$, 并考虑乘积和 $[B_1^{-1}B_0+\Delta t\,C]^n$ 的展开式. 展开式中共有 2^n 项，其中有 $\binom{n}{j}$ 项包含 j 个因子 $\Delta t\,C$ 散布在 $n-j$ 个因子 $B_1^{-1}B_0$ 中；因此后者最多出现 $j+1$ 个顺序相连的因子串，而每一个串的范数的界都是 K_1, 所以每一个这样的项的范数的界是 $K_2^jK_1^{j+1}$. 于是可以得到整个乘积的界

$$\|[B_1^{-1}B_0+\Delta t\,C]^n\| \leqslant \sum_{j=0}^{n}\binom{n}{j}K_1^{j+1}(\Delta t\,K_2)^j \tag{5.129}$$

$$= K_1(1+\Delta t\,K_1K_2)^n\leqslant K_1\mathrm{e}^{n\Delta t\,K_1K_2}, \tag{5.130}$$

而这当 $n\Delta t\leqslant T$ 时是有界的. ■

例如，该引理可以应用于以下情形. 设 $u_t=Lu+a(x)u$, 其中 L 是一个常系数的线性算子，则近似 $u_t=Lu$ 的格式 $B_1U^{n+1}=B_0U^n$ 的稳定性可以用傅里叶方法来分析，且在问题中加上 $a(x)u$ 项后分析结果不变.

第二个原因是许多差分格式是 *耗散的* (dissipative), 这使得傅里叶分析能够应用于变系数问题. 所谓一个差分格式是耗散的是指，存在 $\delta>0$ 和正整数 r，使其增长矩阵的每个特征值都满足形如

$$|\lambda(\mathbf{x},\Delta t,\boldsymbol{\xi})|^2\leqslant 1-\delta|\boldsymbol{\xi}|^{2r} \tag{5.131}$$

的关系式，其中 $\boldsymbol{\xi}=(k_1\Delta x_1,k_2\Delta x_2,\ldots,k_d\Delta x_d)^T$. 这一结果是 Fritz John 揭示的，这也促使

[1] Lax, P.D. and Richtmyer, R.D. (1956), Survey of the stability of linear finite difference equations, *Comm. Pure Appl. Math.* **9**, 267-93.

[2] Strang, G. (1964), Wiener-Hopf difference equations, *J. Math. Mech.* **13**(1), 85-96.

Kreiss 于 1964 年[1], 发展了一套关于双曲型方程类似的理论. 后者的分析是基于 l_2 范数的, 并且大量地应用了能量方法, 也因此给出了诸如 5.9 节中得到的一些很特殊情形的结果.

理论上的最新影响来自于对常微分方程 (ODE) 数值方法的进一步认识, 这种影响将来可能还会更加突出. 我们在定理 5.3 中看到, 当空间算子满足强制性条件时, 由该空间差分算子与 Runge-Kutta 时间步进法相耦合的算法是稳定的. 5.9 节中的例子还显示了这一结果与典型方法的能量分析间的联系. 另一个例子出自于 5.8 节的修正方程分析, 5.8 节的讨论显示了修正方程分析是理解有关算法的很有价值的工具, 但目前还没有由此得到多少严格的结果. 而另一方面, ODE 分析的最新进展却在某些情况下给出了严格的误差界. 其中一个例子是在 Reich [2] 的工作中给出的, 所用的方法在其领域中通常被称为向后误差分析, 他的结果发掘了方法的 *辛*(symplectic) 性质, 这又是 ODE 理论的一个领域, 该领域最近几年取得了十分重大的进展 (有关文献见下面的文献注记). 4.9 节中曾简单地提起过这种概念是如何推广到 PDE 的, 从而给出了 *多辛方法* (multi-symplectic method), 我们期待该方法的优越性会在发展中体现到 PDE 的修正方程分析中.

与对耗散型方法的认识过程类似, 能量方法与傅里叶分析的结合得以发展出一套关于边界条件对于稳定性影响的完整理论. 在第 4 章中, 我们已经提到过 Godunov 和 Ryabenkii 的工作, 他们关于离散边界条件给出了稳定性必要条件的重要判据, 4.12 节中给出了应用该方法的例子. Gustafsson, Kreiss 和 Sundström 的有影响的论文[3] 建立了密切相关的 Lax-Richtmyer 稳定性的充分必要条件; 后续的许多论文导出了有价值的实用判据, 包括严厉的或实用的稳定性所需的判据. 差分格式重要性质之一在该工作中着重讨论的是曾在 4.8 和 4.9 节中提到过的 *群速度* (group velocity) 的稳定性. 将增长因子表示为 $\lambda = \mathrm{e}^{-\mathrm{i}\omega\Delta t}$, 则群速度定义为

$$C(k) := \partial\omega/\partial k. \tag{5.132}$$

推荐读者参考 Trefethen[4] 关于群速度与边界条件稳定性之间关系的精彩论述.

最终, 经过近几年的发展人们对非线性问题有了更为深刻的认识, 其起点是如 (5.20) 那样的基于两个近似解之差的稳定性的定义. 人们早就认识到, 一个格式在真解附近线性化展开的稳定性是其收敛性的必要条件, 但这却远远不足以区分两个非常相似的格式所具有的完全不同的性态, 例如当应用于无粘 Bergers 方程 $u_t + uu_x = 0$ 时. 不过, 通过

---

[1] Kreiss, H.O. (1964), On difference approximations of the dissipative type for hyperbolic differential equations, *Comm. Pure Appl. Math.* **17**, 335-353.

[2] Reich, S. (1999), Backward error analysis for numerical integrators, *SIAM J. Numer. Anal.* **36**, 1549-1570.

[3] Gustafsson, B., Kreiss, H.O. and Sundström, A. (1972), Stability theory of difference approximations for mixed initial boundary value problems. II, *Math. Comp.* **26**(119), 649-686.

[4] Trefethen, L.N. (1982), Group velocity in finite difference schemes, *SIAM Rev.* **24**, 113-136.

引入稳定性阈值 (stability threshold) 的概念， López-Marcos 和 Sanz-Serna[1] 以及其他人建立起了与 Lax 等价定理非常类似的结果. 更具实际重要性的是在 4.7 节介绍的差分格式的总变差不增 (TVD) 性质发展出来的总变差稳定性(TV-stability), 以及满足离散熵条件 (discrete entropy condition) 的格式等概念 (见 LeVeque(1992, 2002)).

文献注记与推荐读物

本章材料的标准参考文献仍然是 Richtmyer 和 Morton (1967), 其中在 Banach 空间的框架下给出了 Lax 等价定理的完整证明. 书中还详细讨论了 Kreiss 和 Buchanan 矩阵定理，以及若干比较简单的稳定性充分条件的判据. 然而这些定理的一些推广和简洁得多的证明出现在最近的文献中，而且这方面的理论已经取得了重大进展，有关这些进展的详细情况，推荐读者参考 5.10 节中列出的原始论文.

年刊系列 *Acta Numerica* 是一个查询有关最新进展综述的方便的资料来源. 例如，在 2003 卷中， Cockburn[2] 和 Tadmor[3] 的两篇论文都与本章所讨论的论题密切相关. 针对 PDE , 辛积分和几何积分方法，以及修正方程分析等的进展及其应用也可查阅这些资料.

习　题

5.1 在均匀网格上考虑对流扩散方程 $u_t + a\,u_x = b\,u_{xx}$, 其中 $b > 0$, 的 Dufort-Frankel 格式

$$U_j^{n+1} = U_j^{n-1} - 2\nu\Delta_{0x}U_j^n + 2\mu(U_{j-1}^n + U_{j+1}^n - U_j^{n+1} - U_j^{n-1}),$$

其中 $\nu = a\Delta t/\Delta x$, $\mu = b\Delta t/(\Delta x)^2$. 证明只需 $\nu^2 \leqslant 1$ 而对 μ 无需任何限制，格式便是稳定的，但为保证其相容性则需对 μ 加以限制.

5.2 在区域 $0 < x < 1$, $t > 0$ 上考虑方程 $u_t = u_{xx}$, 初始条件为 $u(x,0) = f(x)$, 边界条件为 $u_x(0,t) = 0$ 和 $u_x(1,t) + u(1,t) = 0$. 证明当导数边界条件用中心差分格式来近似时，显式方法给出方程组

[1] López-Marcos, J.C. and Sanz-Serna, J.M. (1988), Stability anf convergence in numerical analysis III: linear investigation of nonlinear stability, *IMA J. Numer. Anal.* **8**(1), 71-84.

[2] Cockburn, B. (2003), Continuous dependence and error estimation for viscosity methods, *Acta Numerica*. **12**, 127-180.

[3] Tadmor, E. (2003), Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems, *Acta Numerica* **12**, Cambridge University Press, 451-512.

$$
\begin{aligned}
U_j^{n+1} &= (1-2\mu)U_j^n + \mu(U_{j-1}^n + U_{j+1}^n), \quad j = 1,2,\ldots,J-1,\\
U_0^{n+1} &= (1-2\mu)U_0^n + 2\mu U_1^n,\\
U_J^{n+1} &= (1-2\mu-2\mu\Delta x)U_J^n + 2\mu U_{J-1}^n.
\end{aligned}
$$

证明 $V_j^n = \lambda^n \cos kj\Delta x$ 满足该方程组，只要

$$\lambda = 1 - 4\mu \sin^2 \frac{1}{2}k\Delta x,$$

且 k 满足

$$\Delta x \cot k = \sin k\Delta x.$$

通过作 $\cot k$ 和 $\sin k\Delta x$ 的图，证明以上方程仅有 J 个实根，这些根分别给出不同的网格函数 V_j^n. 验证当 Δx 很小时，方程也有一个复根可近似地表示为 $k\Delta x = \pi + iy\Delta x$, 其中 y 是 $y = \coth y$ 的唯一实根. 证明对于该 k 值，有 $\lambda = 1 - 4\mu\cosh^2 \frac{1}{2}y\Delta x$. 由此推导出虽然对该问题取 $\mu = \frac{1}{2}$ 时，显式方法导致误差增长，但根据判据 (2.55) 和 (2.56)，方法仍然是稳定的.

5.3　考虑二次多项式 $p(z) = az^2 + 2bz + c$, 其中 a, b 和 c 为复数.
(i) 证明如果 $|c| > |a|$，则 $p(z)$ 至少有一个零点的模大于 1.
(ii) 设 $|c| < |a|$, 并定义 $q(z) = \bar{c}z^2 + 2\bar{b}z + \bar{a}$ 和 $r(z) = \bar{a}p(z) - cq(z)$. 证明若 $|z| = 1$ 则 $|q(z)| = |p(z)|$, 因而 $|cq(z)| < |\bar{a}p(z)|$. 由 Rouché 定理推导出 $p(z)$ 和 $r(z)$ 在单位圆内零点的个数相同，因此 $p(z)$ 的两个零点的模均小于或等于 1 的充分必要条件是 $2|\bar{a}b - \bar{b}c| \leqslant |a|^2 - |c|^2$.
(iii) 设 $|a| = |c|$. 记 $c = a\mathrm{e}^{\mathrm{i}\theta}$, $b = a\beta\mathrm{e}^{\mathrm{i}\phi}$, $z = u\mathrm{e}^{\mathrm{i}\theta/2}$, 其中 β 是实数，证明

$$u^2 + 2\beta\mathrm{e}^{\mathrm{i}(\phi-\theta/2)}u + 1 = 0.$$

证明：当且仅当 $\theta = 2\phi$ 且 $|\beta| \leqslant 1$ 时，以上方程的所有根全落在单位圆周上. 证明：若 $|c| = |a|$, 则当且仅当 $\bar{a}b = \bar{b}c$ 且 $|b| \leqslant |a|$ 时，$p(z)$ 的所有零点的模均小于或等于 1.

5.4　在整个实直线上用 MacCormack 格式

$$
\begin{aligned}
U^{n+*} &= U^n - \nu\Delta_- U^n,\\
U^{n+1} &= U^n - \frac{1}{2}\nu(\Delta_- U^n + \Delta_+ U^{n+*}) + \frac{1}{2}\mu(\delta^2 U^n + \delta^2 U^{n+*}),
\end{aligned}
$$

逼近对流扩散方程 $u_t + au_x = bu_{xx}$, 其中 $\nu = a\Delta t/\Delta x$, $\mu = b\Delta t/(\Delta x)^2$ 均为正数. 找到使冯诺伊曼稳定性条件得以满足的条件. 证明：对于实用稳定性，其必要条件是

$$2\mu - 1 \leqslant \nu \leqslant 1,$$

且当 $\mu = \frac{1}{2}$ 时，该条件也是充分的.

5.5　考虑模拟杆振动的差分方程

$$
\begin{aligned}
U^{n+1} - U^n &= -\mu\delta^2 V^n,\\
V^{n+1} - V^n &= \mu\delta^2 U^{n+1}, \qquad \mu = a\Delta t/(\Delta x)^2,
\end{aligned}
$$

找到其稳定性的必要性和充分性判据.

现设杆受到张力作用，在两个方程中分别引入 $\nu\Delta_0 V^n$ 和 $\nu\Delta_0 U^{n+1}$, 其中 $\nu = b\Delta t/\Delta x$. Lax-Richtmyer 稳定性会因此而受到影响吗？找到一个简单的实用稳定性的充分条件.

5.6 设在均匀网格上将 Lax-Wendroff 格式应用于对流方程 $u_t + au_x = 0$, 其中 a 为常数. 记 $\nu = a\Delta t/\Delta x$, 证明：在整个实直线上有

$$\|U^{n+1}\|^2 = \|U^n\|^2 - \frac{1}{2}\nu^2(1-\nu^2)\left[\|\Delta_- U^n\|^2 - \langle\Delta_- U^n,\ \Delta_+ U^n\rangle\right],$$

并由此推出稳定性条件.

如果将该方法用于 (0, 1) 区间，设 $a > 0$, 在 $x = 0$ 处的边界条件为 $U_0^n = 0$, 找到一个简单的使该方法为稳定的 $x = 1$ 处的数值边界条件.

5.7 设将盒式格式应用于方程 $u_t + au_x = 0$, 其中 a 是正常数，区域为 $x > 0$, $t > 0$, 网格是均匀的. 设初始条件为 $u(x,0) = f(x)$, $\forall x > 0$, 其中 $f \in L_2(0,\infty)$; 边界条件为 $u(0,t) = 0$, $\forall t > 0$. 将格式写为

$$U_j^{n+1} + U_{j-1}^{n+1} + \nu(U_j^{n+1} - U_{j-1}^{n+1}) = U_j^n + U_{j-1}^n - \nu(U_j^n - U_{j-1}^n),$$

并定义

$$S_n := \sum_{j=1}^{\infty}\{(U_j^n + U_{j-1}^n)^2 + \nu^2(U_j^n - U_{j-1}^n)^2,$$

其中 $\nu = a\Delta t/\Delta x$, 证明 $S_{n+1} = S_n$. 证明对任意的 Δx 和 Δt ，数值解都是 $l_2(0,\ \infty)$ 有界的.

现设 a 为 x 的正函数，且 $|a'(x)| \leqslant L$, 又设将盒式格式中的 a 换为 $a(x_j - \frac{1}{2}\Delta x)$. 证明

$$S_{n+1} - S_n \leqslant 2L\Delta t\left\{\sum_{j=1}^{\infty}\left[(U_j^n)^2 + (U_j^{n+1})^2\right]\right\}.$$

由此能得出关于该格式稳定性的什么结论？

5.8 设在均匀网格上将 θ- 方法 用于方程 $u_t = u_{xx}$, 设在右端边界给定边界条件 $U_J^n = 0$, 在左端边界给定边界条件 $aU_0^n + bU_1^n = 0$, 其中 a 和 b 为常数. 考虑差分方程在内点上的 $U_j^n = \widehat{U}\lambda^n\mu^j$ 形式的解，找出 λ 和 μ 相互之间应该满足的关系.

证明：如果 $|a/b| \geqslant 1$, 或者 $|a/b| < 1$ 且

$$\frac{2\Delta t(1-2\theta)}{(\Delta x)^2} \leqslant \frac{4ab}{(a+b)^2},$$

则满足 $|\lambda| > 1$ 和 $|\mu| < 1$ 的解必不满足左端的边界条件.

[提示. Godunov-Ryabenkii 的稳定性的必要条件，要求内点方程的任何形如 $\lambda^n\mu^j$ 的解满足：(i) 如果 $|\mu| = 1$, 则 $|\lambda| \leqslant 1$; (ii) 如果 $|\mu| < 1$ 且 $|\lambda| > 1$, 则对应的解必不满足左端的边界条件； (iii) 对 $|\mu| > 1$ 有类似的结果.]

第 6 章　二维线性二阶椭圆型方程

6.1　一个模型问题

同前几章类似，我们将从最简单的模型问题 入手，即求解：

$$u_{xx}+u_{yy}+f(x,y) = 0, \qquad (x,y)\in\Omega, \tag{6.1a}$$

$$u = 0, \qquad (x,y)\in\partial\Omega, \tag{6.1b}$$

其中 Ω 是单位正方形

$$\Omega:=(0,1)\times(0,1), \tag{6.2}$$

而 $\partial\Omega$ 是此正方形的边界. 在第 3 章中，我们曾讨论过以下类型的抛物型方程：

$$\frac{\partial u}{\partial t}=u_{xx}+u_{yy}+f(x,y). \tag{6.3}$$

比较这两个方程，我们发现，如果当 $t\to\infty$ 时，(6.3) 的解收敛，则其极限就是 (6.1a) 的解. 抛物型问题的与时间有关的解和椭圆型问题之间存在的这种关系，经常在椭圆型问题的求解中体现出来. 在第 7 章，我们还将讨论用于求解椭圆型问题的迭代方法和用于求解相应抛物型问题的时间步进有限差分方法之间的联系.

我们在此单位正方形区域上构造正方形一致网格，每个方向上都分为 J 个区间，则

$$\Delta x=\Delta y=1/J. \tag{6.4}$$

我们利用中心差分格式近似 (6.1):

$$\frac{U_{r+1,s}+U_{r-1,s}+U_{r,s+1}+U_{r,s-1}-4U_{r,s}}{(\Delta x)^2}+f_{r,s}=0. \tag{6.5}$$

在 (6.5) 中依次取 $r=1,2,\ldots,J-1$ 和 $s=1,2,\ldots,J-1$，便得到了一个包括 $(J-1)^2$ 个方程的方程组，这个方程组与第 3 章中用隐式方法求解抛物型方程时得到的方程组具有完全相同的结构. 假设已经通过某种方法求得了这些方程的解，我们先来研究结果的精度.

6.2　模型问题的误差分析

如通常做法，我们从截断误差入手. 将微分方程 (6.1) 的精确解 $u(x_r,y_s)$ 代入方程

(6.5), 然后做泰勒级数展开，易知截断误差为

$$T_{r,s} = \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy})_{r,s} + o\left((\Delta x)^2\right). \tag{6.6}$$

截断误差的绝对值有上界 T, 事实上

$$|T_{r,s}| \leqslant T := \frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy}), \tag{6.7}$$

这里采用了 $u(x,y)$ 的偏导数的界的通用记法.

我们在分量为 $U_{r,s}$ 的数组 U 所构成的集合上定义算子 L_h，其在内部节点 $J_\Omega \equiv \{(x_r, y_s); r = 1,2,\ldots,J-1,\ s = 1,2,\ldots,J-1\}$ 上的取值为

$$\begin{aligned}(L_hU)_{r,s} &\equiv L_hU_{r,s} \\ &:= \frac{1}{(\Delta x)^2}(U_{r+1,s} + U_{r-1,s} + U_{r,s+1} + U_{r,s-1} - 4U_{r,s}).\end{aligned} \tag{6.8}$$

这样数值解近似满足方程

$$L_hU_{r,s} + f_{r,s} = 0, \tag{6.9}$$

而精确解则满足方程

$$L_hu_{r,s} + f_{r,s} = T_{r,s}. \tag{6.10}$$

我们按照通常做法定义误差为

$$e_{r,s} := U_{r,s} - u_{r,s}, \tag{6.11}$$

则

$$L_he_{r,s} = -T_{r,s}. \tag{6.12}$$

由于边界各点处的值 $u(x,y)$ 已经给定，边界处的误差值 $e_{r,s}$ 为零.

为给出误差 $e_{r,s}$ 的界，我们首先定义一个**比较函数** (comparison function) :

$$\Phi_{r,s} := \left(x_r - \frac{1}{2}\right)^2 + \left(y_s - \frac{1}{2}\right)^2 \tag{6.13}$$

于是有

$$L_h\Phi_{r,s} = 4. \tag{6.14}$$

直接计算即可验证这个结果，而更容易的做法是，注意到 Φ 是关于 x 和 y 的二次函数，那么 $L_h\Phi$ 即为 $\Phi_{xx} + \Phi_{yy}$ 的精确值，而后者显然等于 4. 若令

$$\psi_{r,s} := e_{r,s} + \frac{1}{4}T\Phi_{r,s}, \tag{6.15}$$

就有

$$
\begin{aligned}
L_h\psi_{r,s} &= L_h e_{r,s} + \frac{1}{4}TL_h\Phi_{r,s} \\
&= -T_{r,s} + T \\
&\geqslant 0, \qquad \forall\,(x_r, y_s) \in J_\Omega.
\end{aligned} \tag{6.16}
$$

以下我们将用到一个最大值原理，它与我们在第 2 章定理 2.2 中用到的类似. 在 6.5 节中，我们将在更一般的意义下给出其证明. 简单地说，算子 L_h 具有这样的性质：如果在某点 (x_r, y_s) 处，有 $L_h\psi_{r,s} \geqslant 0$, 那么 $\psi_{r,s}$ 不可能大于所有邻近点的值. 这样，由 (6.16) 即得， ψ 的正的最大值必然在此正方形区域的边界上取得，而 $e_{r,s}$ 在边界上为零，并且 Φ 在顶角处取得最大值 $\frac{1}{2}$. 从而 ψ 在边界上的最大值是 $\frac{1}{8}T$, 且

$$
\psi_{r,s} \leqslant \frac{1}{8}T \qquad \forall\,(x_r, y_s) \in J_\Omega. \tag{6.17}
$$

注意到 Φ 非负，由 ψ 的定义我们得到

$$
\begin{aligned}
U_{r,s} - u(x_r, y_s) = e_{r,s} &\leqslant \psi_{r,s} \\
&\leqslant \frac{1}{8}T \\
&= \frac{1}{96}(\Delta x)^2(M_{xxxx} + M_{yyyy}).
\end{aligned} \tag{6.18}
$$

以上给出了误差的上界. 定义 $\psi_{r,s} = \frac{1}{4}T\Phi_{r,s} - e_{r,s}$，不难重复以上分析，即可给出误差下界估计 $-e_{r,s} \leqslant \frac{1}{8}T$. 于是得到我们所求的误差界为

$$
|U_{r,s} - u(x_r, y_s)| \leqslant \frac{1}{96}(\Delta x)^2(M_{xxxx} + M_{yyyy}). \tag{6.19}
$$

6.3 一般的扩散问题

现在我们要把这个方法推广至一个更一般的椭圆型问题，即扩散方程

$$
\nabla \cdot (a\nabla u) + f = 0, \qquad \text{在 } \Omega \text{ 内}, \tag{6.20}
$$

其中

$$
a(x, y) \geqslant a_0 > 0. \tag{6.21}
$$

我们假设 Ω 是一个有界开区域，其边界为 $\partial\Omega$. 边界条件可能具有一般的形式：

$$
\alpha_0 u + \alpha_1 \partial u/\partial n = g, \qquad \text{在 } \partial\Omega \text{ 上}, \tag{6.22}
$$

这里 $\partial/\partial n$ 表示求外法向上的导数，并且

$$
\alpha_0 \geqslant 0, \quad \alpha_1 \geqslant 0, \quad \alpha_0 + \alpha_1 > 0. \tag{6.23}
$$

与上一章类似，我们在区域 Ω 上构造正则网格 (regular mesh), 在 x 方向的尺寸为 Δx, 在 y 方向的尺寸为 Δy.

假设 a 光滑变化，并记 $b = \partial a/\partial x$, $c = \partial a/\partial y$. 我们可以将 (6.20) 展开为

$$a\nabla^2 u + bu_x + cu_y + f = 0. \tag{6.24}$$

在离开边界的点处，我们可以用中心差分格式近似这个方程，得到的近似值 $U := \{U_{r,s},\ (r,s) \in J_\Omega\}$ 满足

$$a_{r,s}\left[\frac{\delta_x^2 U_{r,s}}{(\Delta x)^2} + \frac{\delta_y^2 U_{r,s}}{(\Delta y)^2}\right] + b_{r,s}\left[\frac{\Delta_{0x} U_{r,s}}{\Delta x}\right] + c_{r,s}\left[\frac{\Delta_{0y} U_{r,s}}{\Delta y}\right] + f_{r,s} = 0. \tag{6.25}$$

我们采用通常方法定义这个五点格式的截断误差，容易发现，此误差关于 Δx 和 Δy 都是二阶的. 在此方程中含有 $U_{r+1,s}$ 和 $U_{r-1,s}$ 的项是

$$\left[\frac{a_{r,s}}{(\Delta x)^2} - \frac{b_{r,s}}{2\Delta x}\right] U_{r-1,s} + \left[\frac{a_{r,s}}{(\Delta x)^2} + \frac{b_{r,s}}{2\Delta x}\right] U_{r+1,s}. \tag{6.26}$$

为了用最大值原理来分析该格式的误差，就必须确保与 (r,s) 点相邻的各点处的所有系数都有相同的正负号. 显然，这就要求

$$|b_{r,s}|\Delta x \leqslant 2a_{r,s} \qquad \forall r, s \tag{6.27}$$

对 $c_{r,s}$ 也有类似的限制. 这就意味着需要在扩散系数 $a(x,y)$ 较小，变化较快的区域采用较密的网格.

不过，正如我们在 2.8 节考虑极坐标时所看到的，一个更自然的格式则更直接地基于 (6.20) 式的积分形式. 考虑一个网格节点周围的控制体 V, 在图 6-1 中，我们用点线表示这个在上一章引入的概念. 在此控制体上积分 (6.20), 利用高斯定理，我们得到

$$\int_{\partial V} a(\partial u/\partial n)\mathrm{d}l + \int_V f\mathrm{d}x\mathrm{d}y = 0. \tag{6.28}$$

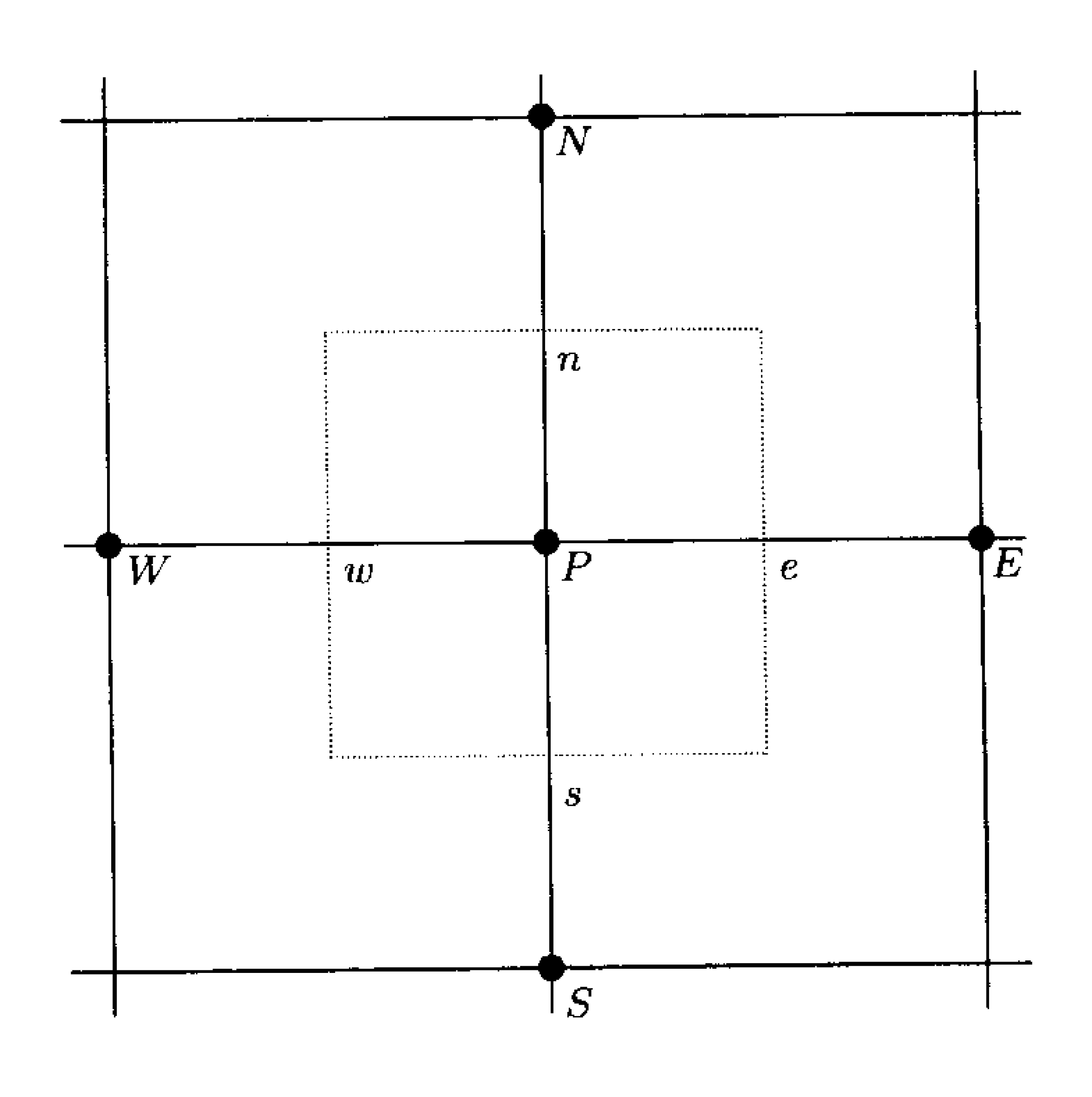


图 6-1　点 P 所在的控制体

现在我们可以做 (6.28) 各项的近似以构造一个差分格式. V 的边界 ∂V 的法向都在坐标轴的方向上, 可以利用与 (6.25) 式中相同的 5 个点, 以其差商来近似法向导数. 用这个矩形的 4 条边中每边的长度和该边中点的法向导数值的乘积, 来近似这条边上的线积分. 同样地, 我们用这个单元的面积和在此单元中点处的值的乘积, 来近似 $f(x,y)$ 在这个单元上的积分. 这样做的结果是, 我们得到了格式

$$\begin{aligned}&\frac{\Delta y}{\Delta x}\left[a_{r+1/2,s}(U_{r+1,s}-U_{r,s})-a_{r-1/2,s}(U_{r,s}-U_{r-1,s})\right]\\&\quad+\frac{\Delta x}{\Delta y}\left[a_{r,s+1/2}(U_{r,s+1}-U_{r,s})-a_{r,s-1/2}(U_{r,s}-U_{r,s-1})\right]+\Delta x\Delta y f_{r,s}=0\end{aligned} \tag{6.29a}$$

或写作

$$\left[\frac{\delta_x(a\,\delta_x U)}{(\Delta x)^2}+\frac{\delta_y(a\,\delta_y U)}{(\Delta y)^2}\right]_{r,s}+f_{r,s}=0\,. \tag{6.29b}$$

我们可以采用如图 6-1 所示的“罗经点” 表示法, 并将 (6.29a) 、 (6.29b) 写作

$$\frac{a_e(U_E-U_P)-a_w(U_P-U_W)}{(\Delta x)^2}+\frac{a_n(U_N-U_P)-a_s(U_P-U_S)}{(\Delta y)^2}+f_P=0. \tag{6.30}$$

这样做常常带来方便. 由于我们已经假设函数 $a(x,y)$ 总是正的, 因此容易看出格式 (6.30) 中的系数总有正确的正负号, 而不需要对网格的大小做出限制.

在一类问题中, 存在一个材料性质 (用 a 来表示) 发生骤变, 但是法向通量 $a\partial u/\partial n$ 保持连续的界面. 在处理这种问题时, (6.29) 形式有其优越性. 假设我们适当地构造网格, 使得材料界面垂直穿过点 e, 在界面右边保持常量 a_E, 在左边保持 a_P. 这样利用界面处的中间值 U_e, 在界面处连续的通量值可以近似为, 例如

$$\begin{aligned}\frac{a_E(U_E-U_e)}{\frac{1}{2}\Delta x}&=\frac{a_P(U_e-U_P)}{\frac{1}{2}\Delta x}\\&=\frac{a_e(U_E-U_P)}{\Delta x},\end{aligned} \tag{6.31}$$

这里我们通过定义 a_e:

$$\frac{2}{a_e}=\frac{1}{a_E}+\frac{1}{a_P}. \tag{6.32}$$

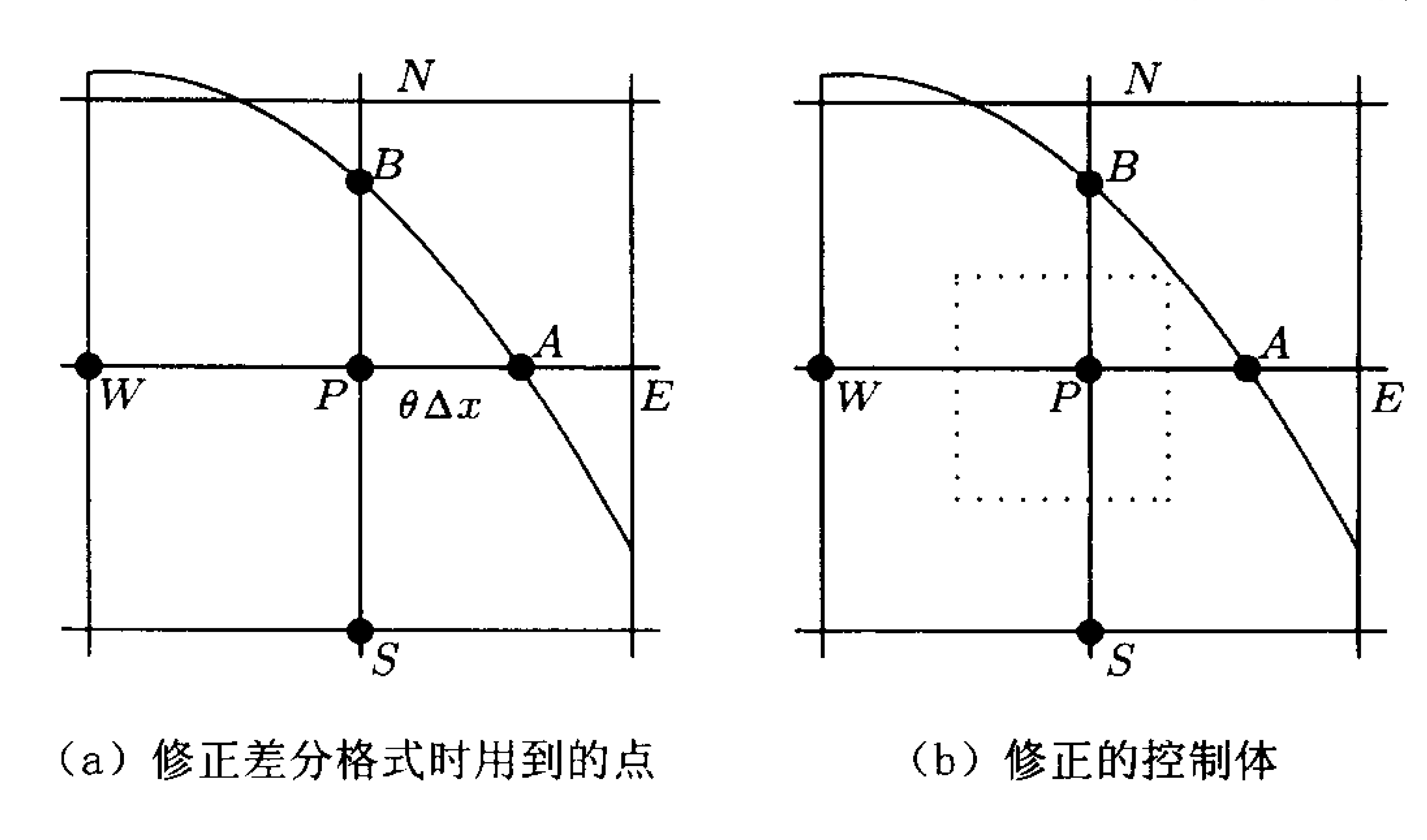
消去 U_e, 得到最后一个表达式, 只待代入 (6.30).

6.4 曲线边界上的边值条件

格式 (6.25) 或者 (6.30) 需要在不顺着网格线的边界的邻域处做修正. 在 3.4 节, 我们了解到, 在这些点处是如何近似二阶导数的, 现在我们必须把这种方法推广到更一般的

扩散方程.

我们首先考虑如图 6-2 所示情况，在曲线边界上给出狄利克雷条件，如果我们利用



（a）修正差分格式时用到的点　　（b）修正的控制体

图 6-2　在曲线边界上的狄利克雷条件

泰勒级数展开来获得 (6.24) 在点 P 处形如 (6.25) 的近似，我们需要用到点 A 和 B 处而不是点 E 和 N 处给定的值. 这里只考虑点 A 即可，我们记 $PA=\theta\Delta x$, 那么 u_A 和 u_W 的泰勒级数展开就是

$$\begin{aligned} u_A &= [u+\theta\Delta x\, u_x+\frac{1}{2}(\theta\Delta x)^2 u_{xx}+\cdots]_P, \\ u_W &= [u-\Delta x\, u_x+\frac{1}{2}(\Delta x)^2 u_{xx}-\cdots]_P. \end{aligned}$$

由此，先后消去 u_{xx} 和 u_x, 我们获得如下近似：

$$[u_x]_P \approx \frac{u_A-\theta^2 u_W-(1-\theta^2)u_P}{\theta(1+\theta)\Delta x}, \tag{6.33a}$$

$$[u_{xx}]_P \approx \frac{u_A+\theta u_W-(1+\theta)u_P}{\frac{1}{2}\theta(1+\theta)(\Delta x)^2}. \tag{6.33b}$$

对 U_P, U_S 和 U_B 做相同地处理，我们便得到 (6.25) 的一个适当的修正格式. 与在通常内点处相同，为使最大值原理成立，我们仍需对网格的尺寸做出限制.

差分格式 (6.30) 的积分形式在狄利克雷边界附近可做类似地修正. 如图 6-1, 在每个网格点周围我们做一矩形控制体，但是，如图 6-2(b), 如果连接点 P 和它的一个邻近节点的网格线段与边界相交，那么，矩形控制体相应的边要与该网格线垂直，且过连接 P 和其邻近节点的线段的中点. 例如，在这幅图中，距离 PA 是网格尺寸 $PE=\Delta x$ 的一部分 θ, 所以控制体 V 的宽度是 $\frac{1}{2}(1+\theta)\Delta x$.

这样，沿着这个单元底边的线积分近似为

$$\frac{1}{2}(1+\theta)\Delta x\left[\frac{-a_s(U_P-U_S)}{\Delta y}\right]. \tag{6.34}$$

我们还必须将沿着单元右边侧的线积分中的法向导数近似调整为

$$\frac{a_a(U_A - U_P)}{\theta \Delta x}, \tag{6.35}$$

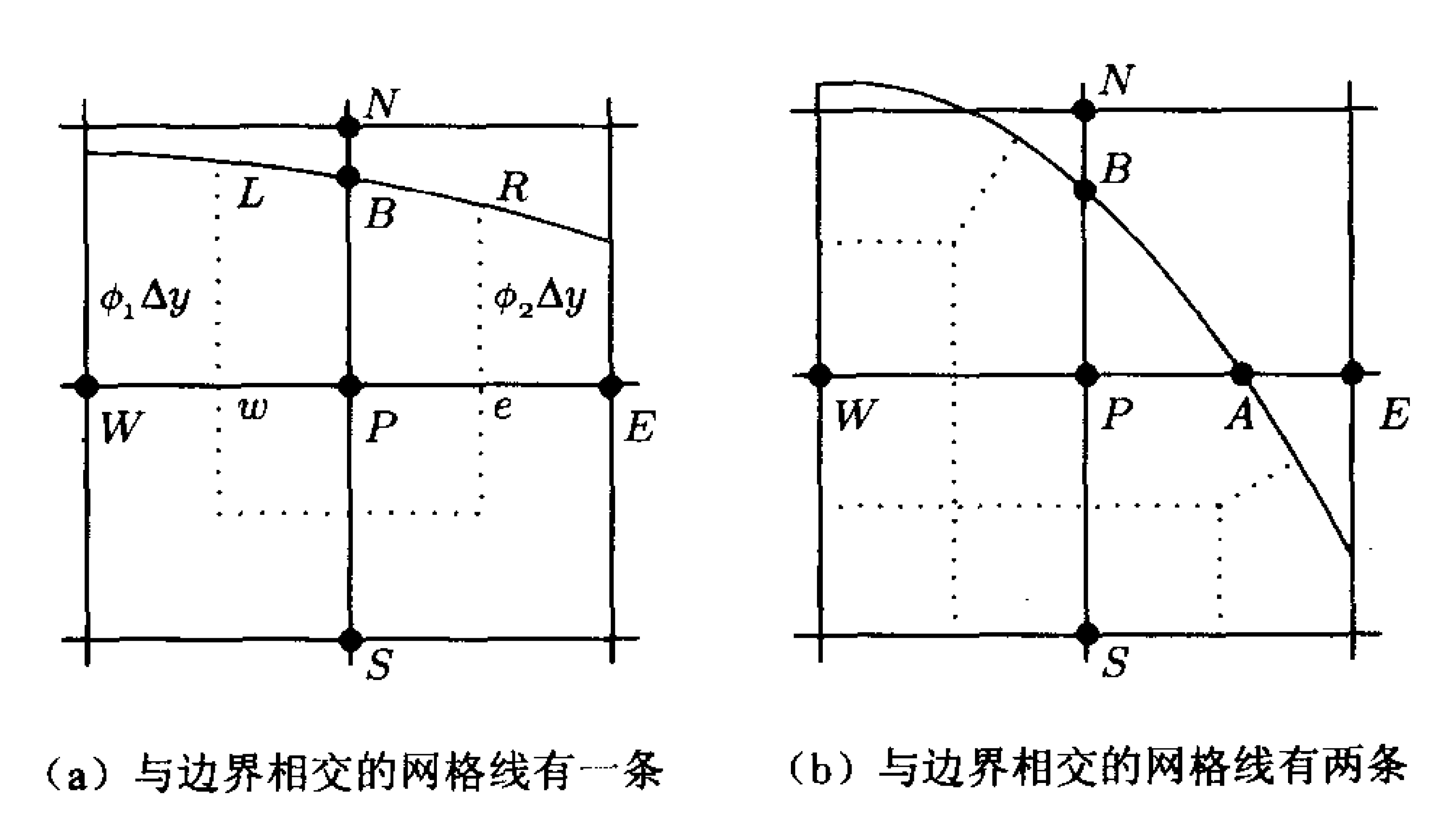
其中，a_a 是 $a(x,y)$ 在 A 和 P 的中间点处的值，而 a_b 也有类似的意义. 注意到在图 6-2(b) 中，边界与网格线相交于两点 A 和 B, 且 $PA = \theta\Delta x$, $PB = \phi\Delta y$, 于是在 P 点处我们得到的差分逼近为:

$$\begin{aligned}\frac{1}{2}(1+\phi)\Delta y &\left[\frac{a_a(U_A - U_P)}{\theta\Delta x} - \frac{a_w(U_P - U_W)}{\Delta x}\right] \\ &+ \frac{1}{2}(1+\theta)\Delta x\left[\frac{a_b(U_B - U_P)}{\phi\Delta y} - \frac{a_s(U_P - U_S)}{\Delta y}\right] \\ &+ \frac{1}{4}(1+\theta)(1+\phi)\Delta x\Delta y f_P = 0. \end{aligned} \tag{6.36}$$

显然，当 $a(x,y)$ 是常量时，这个格式与由 (6.33a, b) 给出的格式是一样的. 在更一般的情况下，这个格式的好处就是其系数仍然满足最大值原理所要求的条件.

导数型边界条件更难处理. 如我们在第 3 章中看到的，对法向导数难以构造精确的差分逼近，而且有必要考虑很多不同的可能的几何构型. 另外，在 6.7 节中，我们将看到，用有限元逼近导数型边界条件更为直接. 不过，我们将说明如何在一种简单的情况下处理积分形式的方程.

考虑如图 6-3(a) 所示 的边界与网格线 PN 交于点 B，但与其他连接 P 点与其紧邻节点的连线不相交. 我们还假设在这部分边界上给定诺伊曼条件



（a）与边界相交的网格线有一条 （b）与边界相交的网格线有两条

图 6-3 诺伊曼条件的积分形式

$$\frac{\partial u}{\partial n} = g, \tag{6.37}$$

形如 (6.22) 的一般边界条件则更难处理. 我们做一控制体 V 包围 P 点，先按照以前做法画出一个矩形的三条边，然后延长两条竖直边使之与边界相交于点 L 和 R, 利用边界在

点 L 和 R 之间的部分补全控制体边界. 记

$$wL = \phi_1 \Delta y, \qquad eR = \phi_2 \Delta y \tag{6.38}$$

我们可以完全按照以前的做法近似水平和竖直边上的线积分, 但是, 注意, 由于我们计算的不是竖直边上中点处的法向导数, 结果的精度将会降低. 同样, 二重积分由单元的面积和 f 在 P 处的值的乘积近似. 由于 P 不在单元的中心, 这里又出现了精度损失.

最后我们必须近似沿边界 RBL 的线积分, 这里记

$$\int a \frac{\partial u}{\partial n}\, \mathrm{d}l = \int ag\, \mathrm{d}l \tag{6.39}$$

后者可以近似为

$$a_B\, g_B\, \psi\, \Delta x \tag{6.40}$$

其中 $\psi \Delta x$ 是线段 LR 的长度. 最终导出了差分方程

$$\left(\phi_2 + \frac{1}{2}\right) \Delta y \left[a_e \frac{U_E - U_P}{\Delta x}\right] + \left(\phi_1 + \frac{1}{2}\right) \Delta y \left[a_w \frac{U_W - U_P}{\Delta x}\right] + \Delta x \left[a_s \frac{U_S - U_P}{\Delta y}\right] + \psi \Delta x\, a_B\, g_B + \frac{1}{2}(\phi_1 + \phi_2 + 1)\Delta x \Delta y f_P = 0. \tag{6.41}$$

在边界与两条 P 与其邻近点的连线相交这种很普遍的情况下, 处理步骤几乎一样, 不过, 如图 6-3(b) 所示, 构造单元时必须用到某些对角线, 其几何上的细节非常繁琐, 这里我们就不再进一步讨论了.

6.5 利用最大值原理的误差分析

假设在近似 (6.20), (6.22), 或者任何线性椭圆型方程时, 在每一个内点 $P \in J_\Omega$ 处, 都构造了一个如下形式的近似

$$L_h U_P + f_P + g_P = 0, \tag{6.42}$$

这里 g_P 表示任何由非狄利克雷边值条件给出的边值信息. 我们假设以下条件成立:

(i) 对每个 $P \in J_\Omega$, L_h 具有形式

$$L_h U_P = \sum_k c_k U_k - c_P U_P, \tag{6.43}$$

这里系数都是正的, 并且关于 k 的求和是针对与 P 相邻的网格点的. 在上述差分格式中, 这些点仅包括四个最近的邻点, 但是这种分析可以应用于涵盖了诸如对角线方向上的 U_{NE} 的更一般的差分格式. 还有, 当 P 在边界附近时, 某些邻点, 如图 6-2 和方程 (6.36) 中的 A 和 B, 有可能成为边界上的点集 $J_{\partial\Omega}$ 中的元素, 并且, 相应的值

U_k 由狄利克雷边界条件给出. 关键的要求是, 在格式 (6.43) 中出现的所有系数都必须是正的.

(ii) 对每个 $P \in J_\Omega$, 都有

$$c_P \geqslant \sum_k c_k. \tag{6.44}$$

(iii) 集合 J_Ω 是**连通的** (connected). 我们说点 P 与每一个出现在 (6.43) 中的具有非零系数的邻点是**连通的**, 如果给出集合 Ω_J 中的任意两点 P 和 Q, 总存在一列点 $P = P_0, P_1, P_2, \ldots, P_m = Q$, 使得对 $r = 1, 2, \ldots, m-1$, 每个点 P_r 都与 P_{r-1} 和 P_{r+1} 连通, 那么这个集合就是连通的.

(iv) (6.43) 诸方程中必须至少有一个包含一个由狄利克雷边界条件给出的边界值. 换言之, 必须在至少一部分边界上赋狄利克雷边界条件. (为了保证 (6.1) 的解的唯一性需要这样一个条件, 或者至少需要 $\alpha_0 \neq 0$ 时的边界条件 (6.22).)

要注意, 在这些条件中, 我们是在一个特定的意义下引证 内(interior) 点 J_Ω 和 边界(boundary) 点 $J_{\partial\Omega}$ 的. 在一些点处, (6.42) 形式的方程成立, 其系数满足适当的条件, 内点就是指这些点. 另一些点也出现在至少一个方程中, 但在这些点处, 没有给出任何方程, 而其 U 的值是通过狄利克雷边界条件预先指定的, 边界点就是指这些点. 在此区域内给定了诺伊曼边界条件或者混合边界条件的边界点处, 我们一般要消去 U 的值, 如同图 6-3 所示的那种情况, 这时未知的值 U_B 并没有出现在关于 U_P 的方程 (6.41) 或任何其他方程中; 或者我们视其为一个拥有关于它自己的方程的内点, 如在下面 6.9 节例子中对对称边界条件的处理. 由此 边界(boundary) 点即为由狄利克雷条件预先给定的那些点. 注意, 这些定义和约定就我们现在的目的来说很方便, 但与我们在第 5 章中用到的以及将在第 7 章使用的略有不同.

引理 6.1. (最大值原理) *假设 L_h, J_Ω 和 $J_{\partial\Omega}$ 满足上述所有假设, 且网格函数 U 满足*

$$L_h U_P \geqslant 0, \qquad \forall P \in J_\Omega. \tag{6.45}$$

那么 U 不可能在任何内点取到非负的最大值, 即

$$\max_{P \in J_\Omega} U_P \leqslant \max\left\{ \max_{A \in J_{\partial\Omega}} U_A,\ 0 \right\}. \tag{6.46}$$

证明. 我们用反证法. 假设在点 P 处取到内部的最大值 $M_\Omega \geqslant 0$, 且 $M_\Omega > M_{\partial\Omega}$, 后者为边界点上的最大值. 则由 (6.45), (6.43) 和 (6.44), 有

$$M_\Omega = U_P \leqslant \frac{1}{c_P}\sum_k c_k U_k \\ \leqslant \frac{1}{c_P}\sum_k c_k M_\Omega \leqslant M_\Omega. \qquad (6.47)$$

从而 (6.47) 各项均相等，这意味着涉及到的所有的值 U_k 均与 M_Ω 相等. 这样最大值也可以在 P 的与其连通的邻点上取得，对这些点中的每一个都可以重复相同的论证. 由于我们已经假设 J_Ω 是连通的，这就表明 U 在所有的内点处都取得相同的值 M_Ω, 并且这些点中至少有一个在边界上的连通邻点，这与 $M_\Omega > M_{\partial\Omega}$ 的假设矛盾. ■

推论. 若

$$L_h U_P \leqslant 0, \qquad \forall P \in J_\Omega$$

则

$$\min_{P\in J_\Omega} U_P \geqslant \min\left\{\min_{A\in J_{\partial\Omega}} U_A, 0\right\}. \qquad (6.48)$$

证明的方法与前一个证明相同，或者我们只要在网格函数 U 上应用引理即可.

我们接下来用通常的做法在每个内点处定义截断误差，注意到 (6.42) 已经做了适当的比例变换，

$$T_P := L_h u_P + f_P + g_P. \qquad (6.49)$$

由此，对每个内点处的误差 $e_P = U_P - u_P$, 我们有

$$L_h e_P = -T_P, \qquad P \in J_\Omega. \qquad (6.50)$$

如通常的做法，我们将假设在所有给定了狄利克雷边值的边界点处 $e_A = 0$, 即狄利克雷条件是精确的强制性条件，这就使我们可以利用引理 6.1 给出的最大值原理界定 e_P.

定理 6.1. 假设定义在 $J_\Omega \cup J_{\partial\Omega}$ 上的非负网格函数 Φ 满足

$$L_h \Phi_P \geqslant 1, \qquad \forall\ P \in J_\Omega, \qquad (6.51)$$

和前面的四个条件，则格式 (6.42) 的误差界由下式给出：

$$|e_P| \leqslant \left[\max_{A\in J_{\partial\Omega}} \Phi_A\right]\left[\max_{P\in J_\Omega} |T_P|\right]. \qquad (6.52)$$

证明. 我们用 T 表示截断误差的绝对值 $|T_P|$ 的最大值. 因为

$$L_h(T\Phi_P + e_P) \geqslant T - T_P \geqslant 0, \qquad (6.53)$$

所以我们可以对函数 $T\Phi_P + e_P$ 应用引理 6.1, 再利用 Φ 非负的事实，我们可由此得到

$$\max_{P\in J_\Omega} e_P \leqslant \max_{P\in J_\Omega}(T\Phi_P + e_P) \leqslant \max_{A\in J_{\partial\Omega}}(T\Phi_A + e_A),$$

即

$$\max_{P\in J_\Omega} e_P \leqslant \left[\max_{A\in J_{\partial\Omega}} \Phi_A\right] T, \tag{6.54}$$

这里我们利用了这样的事实：因为我们假设采用了给定的狄利克雷边值条件的精确值，所以在所有的边界点处 $e_A=0$.

以上给出了误差上界. 如果我们对函数 $-e_P$ 应用引理，我们将对 $T\Phi_P-e_P$ 得到类似的界限，继而得到需要的结果. ■

这里的陈述和分析都是在相当一般的条件下给出的， 6.2 节中模型问题的分析方法也如出一辙，唯一的区别在于所用到的是一个满足 $L_h\Phi_P=4$ 的比较函数 Φ. 后面将会给出一个涵盖了这两种情形的更一般的定理. 在应用这种方法处理某个具体问题时，我们必须给截断误差找到一个界 T, 并依此构造函数 Φ. 显然，只需要用泰勒级数展开即可非常直接地确定 T, 而构造一个合适的 Φ 可能就会困难些，当然，这个函数并不唯一. 譬如说，在模型问题中，我们就可以定义

$$\Phi_{r,s}=\frac{1}{4}\left[(x_r-p)^2+(y_s-q)^2\right]; \tag{6.55}$$

对任意常数 p 和 q, 所需的条件都能得到满足，而前面所选的 $p=q=\frac{1}{2}$ 是使 $\max(\Phi_A)$ 取得最小值的特殊值.

现在，我们要应用这个一般的方法来分析一些更为复杂的问题. 首先考虑在 6.4 节中讨论过的曲边区域上的泊松方程 (6.1) 的解， 其有限差分近似满足最大值原理，并且在其全部四个邻点都是内部网格点的点处，截断误差都具有 (6.6) 的形式，并且满足 (6.7). 不过，对于那些挨着边界的点，也就是有一个或多个邻点在边界上的点，我们必须用更一般的差分近似，例如 (6.33b). 利用泰勒级数展开易得

$$\frac{u_A+\theta u_W-(1+\theta)u_P}{\frac{1}{2}\theta(1+\theta)(\Delta x)^2}=(u_{xx})_P-\frac{1}{3}(1-\theta)\Delta x(u_{xxx})_P+O\left((\Delta x)^2\right). \tag{6.56}$$

一个内部网格点有可能有多个邻点在边界上，但是，由于 (6.56) 中 $0<\theta<1$, 容易看出，在所有情况下，我们总能够选择正数 K_1 和 K_2, 使得只要 Δx 足够小，就有

$$|T_{r,s}|\leqslant K_1(\Delta x)^2 \quad \text{在普通点处,} \tag{6.57a}$$

$$|T_{r,s}|\leqslant K_2\Delta x \quad \text{邻近边界点处.} \tag{6.57b}$$

因此，在所有内部网格点处，都有

$$|T_{r,s}|\leqslant K_1(\Delta x)^2+K_2\Delta x. \tag{6.58}$$

现在假设整个区域是包含在一个以 (p,q) 为圆心， R 为半径的圆中，并按照 (6.55) 定义比较函数 $\Phi_{r,s}$. 则同前， $L_h\Phi_P=1$ 在所有普通的内部网格点处成立. 这个结果也在挨

着边界的点处成立，这是由于 (6.56) 中的截断误差含有三阶和四阶导数，因而对二次多项式截断误差为零. 这样我们可以应用定理 6.1, 由于在整个区域上以及边界处 $0 \leqslant \Phi \leqslant \frac{1}{4}R^2$, 我们推出

$$|U_{r,s} - u(x_r, y_s)| \leqslant \frac{1}{4}R^2[K_1(\Delta x)^2 + K_2\Delta x], \tag{6.59}$$

这说明当网格尺寸趋于零时，误差是 $O(\Delta x)$, 而不是简单区域上的 $O((\Delta x)^2)$.

但是，我们可以利用一个稍作修改的比较函数求得更精确的误差界. 在分析这类问题时，通常的做法是，截断误差在普通的网格点处取某种形式，而在邻近边界点处取另一种形式. 因此，定理 6.1 的一个推广形式会给应用带来方便.

定理 6.2. *我们仍沿用定理 6.1 的记法. 假设集合 J_Ω 分解为两个不相交的子集*

$$J_\Omega = J_1 \cup J_2, \qquad J_1 \cap J_2 = \varnothing;$$

在 $J_\Omega \cup J_{\partial\Omega}$ 上定义了非负网格函数 Φ, 满足

$$\begin{aligned} L_h\Phi_P &\geqslant C_1 > 0, \qquad \forall P \in J_1, \\ L_h\Phi_P &\geqslant C_2 > 0, \qquad \forall P \in J_2; \end{aligned} \tag{6.60}$$

并且差分逼近 (6.42) 的截断误差满足

$$\begin{aligned} |T_P| &\leqslant T_1, \qquad \forall P \in J_1, \\ |T_P| &\leqslant T_2, \qquad \forall P \in J_2. \end{aligned} \tag{6.61}$$

则逼近误差有如下的界

$$|e_P| \leqslant \left[\max_{A \in J_{\partial\Omega}} \Phi_A\right] \max\left\{\frac{T_1}{C_1}, \frac{T_2}{C_2}\right\}. \tag{6.62}$$

证明. 证明是定理 6.1 证明的简单推广. 只需要对函数 $K\Phi + e$ 应用引理 6.1, 这里应适当选择 K 以确保可以应用最大值原理. 证明的细节留作练习. ■

我们对具有曲线边界的问题应用此定理. 取集合 J_1 包含所有普通内部网格点，J_2 包含所有有一个或多个邻点在边界上的网格点. 定义如下的网格函数 Φ

$$\begin{aligned} \Phi_P &= E_1\{(x_r - p)^2 + (y_s - q)^2\} \qquad &\forall P \in J_\Omega, \\ \Phi_P &= E_1\{(x_r - p)^2 + (y_s - q)^2\} + E_2 \qquad &\forall P \in J_{\partial\Omega}, \end{aligned}$$

这里 E_1 和 E_2 是待定的正常数，则

$$L_h\Phi_P = 4E_1, \qquad \forall P \in J_1. \tag{6.63a}$$

但对 J_2 中的点，还有来自边界点的一个或多个附加项. 在逼近式 (6.33b) 中，u_A 的系数为

$$\frac{2}{\theta(1+\theta)(\Delta x)^2}.$$

由于 $0<\theta<1$, 该系数具有不小于 $1/(\Delta x)^2$ 的正下界．因而

$$L_h\Phi_P \geqslant 4E_1 + E_2/(\Delta x)^2 \geqslant E_2/(\Delta x)^2, \qquad \forall\, P \in J_2. \tag{6.63b}$$

令 (6.61) 的截断误差界由 (6.57) 给出，其中取 $T_1 = K_1(\Delta x)^2$ 及 $T_2 = K_2\Delta x$, 则应用定理 6.2, 我们便得到

$$|e_P| \leqslant (E_1R^2 + E_2)\max\left\{\frac{K_1(\Delta x)^2}{4E_1}, \frac{K_2(\Delta x)^3}{E_2}\right\}. \tag{6.64}$$

这个界仅与比例 E_2/E_1 有关，并且当 $\max\{\cdot,\cdot\}$ 中的两个量相等时取到最优值．于是我们有结果

$$|e_P| \leqslant \frac{1}{4}K_1R^2(\Delta x)^2 + K_2(\Delta x)^3, \tag{6.65}$$

这说明误差关于网格尺寸实际上是二阶的．注意，该误差界的主项并没有受到边界邻近点处截断误差中低阶项的影响．

作为第二个例子，我们考虑单位正方形上的泊松方程，在右边界 $x=1$ 上，赋诺伊曼边界条件 $u_x(1,y)=g(y)$, 而在其他三条边上赋狄利克雷条件．如第 2 章中的类似问题，我们在边界 $x=1$ 外的一条线上引进额外的一列点，且 $r=J+1$. 边界条件由下式近似

$$\frac{U_{J+1,s} - U_{J-1,s}}{2\Delta x} = g_s. \tag{6.66}$$

我们从 $r=J$ 处的标准差分方程中消去额外的未知量 $U_{J+1,s}$, 得

$$\frac{U_{J,s+1} + U_{J,s-1} + 2U_{J-1,s} - 4U_{J,s} + 2g_s\Delta x}{(\Delta x)^2} + f_{J,s} = 0. \tag{6.67}$$

这个方程具有了 (6.42) 所示的一般形式，满足最大值原理所要求的条件，所以，在应用最大值原理时，这些 $r=J$ 的点可视为内(internal) 点处理．

普通点处的截断误差同以前一样处理， (6.67) 的截断误差由泰勒级数展开给出，其结果为

$$\begin{aligned} T_{r,s} &= \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy}) + O((\Delta x)^4), \qquad r<J, \\ T_{J,s} &= \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy}) - \frac{1}{3}\Delta x\, u_{xxx} + O((\Delta x)^3). \end{aligned} \tag{6.68}$$

利用定理 6.1 和 (6.13) 给出的比较函数 Φ, 应用前面同样的论证方法便证明了误差界为

$$|e_{r,s}| \leqslant \frac{1}{8}\left\{\frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy}) + \frac{1}{3}\Delta x M_{xxx}\right\}. \tag{6.69}$$

此误差界的阶是 $O(\Delta x)$, 但是如同上一个例子，选取不同的比较函数，应用定理 6.2, 可以得到一个更精确的误差界．定义

$$\Phi = (x-p)^2 + (y-q)^2, \tag{6.70}$$

其中 p 和 q 为待定系数. 我们把区域的内点分为 J_1 和 J_2, 其中 J_2 由 $r = J$ 的点组成. 在区域 J_1 中采用标准的差分方程，则有

$$L_h\Phi_p = 4, \qquad P \in J_1. \tag{6.71}$$

在 $r = J$ 的点处，则采用一个不同的差分算子，如在 (6.67) 中一样，于是有

$$L_h\Phi_P = 4 - 4(1-p)/\Delta x, \quad P \in J_2. \tag{6.72}$$

写成定理 6.2 的记法，我们就会发现

$$\begin{aligned} \frac{T_1}{C_1} &= \frac{\frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy})}{4}, \\ \frac{T_2}{C_2} &= \frac{\frac{1}{12}(\Delta x)^2(M_{xxxx} + M_{yyyy}) + \frac{1}{3}\Delta x M_{xxx}}{4 - (1-p)/\Delta x}. \end{aligned} \tag{6.73}$$

如果我们选取 $p = 2, q = \frac{1}{2}$, 便得到

$$\frac{T_2}{C_2} \leqslant \frac{1}{3}(\Delta x)^2 M_{xxx} + \frac{1}{12}(\Delta x)^3(M_{xxxx} + M_{yyyy}), \tag{6.74}$$

并且在正方形的所有点处有 $(x-2)^2 + (y-\frac{1}{2})^2 \leqslant \frac{17}{4}$; 所以将 T_1/C_1 和 T_2/C_2 相加，我们得到误差界为

$$|e_{r,s}| \leqslant \frac{17}{4}(\Delta x)^2\left[\frac{1}{3}M_{xxx} + \frac{1}{12}(\frac{1}{4} + \Delta x)(M_{xxxx} + M_{yyyy})\right]. \tag{6.75}$$

这说明，在这个例子中，误差关于网格尺寸也是二阶的.

同样的技巧可以用来证明对建立在相当一般的区域上的，在曲线边界上给定了狄利克雷边界条件或诺伊曼边界条件的泊松方程，我们的近似解仍然具有二阶精度. 只是由于需要考虑各种不同的可能的几何形状，使其应用变得复杂化. 事实上，相同的思想可以相当普遍地应用于最大值原理成立的椭圆型和抛物型问题. 在这一节的最后，我们来改进在第 2 章中对一维热流问题给出的一些结果.

在 2.11 节，我们给出了一个关于热方程的 θ- 方法的收敛性的证明. 现在我们这样定义算子 L_h

$$L_h\psi = \frac{\theta\delta_x^2\psi_j^{n+1} + (1-\theta)\delta_x^2\psi_j^n}{(\Delta x)^2} - \frac{\psi_j^{n+1} - \psi_j^n}{\Delta t}. \tag{6.76}$$

并取 P 为点 (x_j, t_{n+1}). 容易看出，只要 $0 \leqslant \theta \leqslant 1$ 且 $\mu(1-\theta) \leqslant \frac{1}{2}$, 则定理 6.1 的条件对一个适当定义的点集 J_Ω 是满足的. 我们假设在 $x = 0$ 和 $x = 1$ 处给定狄利克雷条件，那么内点集 J_Ω 就是那些 $1 \leqslant j \leqslant J-1$ 和 $1 \leqslant n \leqslant N$ 的点，而对边界点 $J_{\partial\Omega}$, 则有 $j = 0$ 或 $j = J$ 或 $n = 0$. 按照 2.10 节的记法，精确解 u 和数值解 U 满足 $L_h u_j^n = -T_j^{n+1/2}$ 和 $L_h U_j^n = 0$, 所以

$$L_h e_j^n = T_j^{n+1/2},$$

且有

$$T_j^{n+1/2} = O(\Delta t) + O((\Delta x)^2).$$

现在，定义比较函数

$$\Phi_j^n = At_n + Bx_j(1-x_j), \tag{6.77}$$

这里 A 和 B 是非负常数；容易证明 $L_h\Phi_j^n = -(A+2B)$. 所以，如果选取各常数使 $A+2B \geqslant T := \max|T_j^{n+1/2}|$ 成立，则有

$$L_h(e_j^n - \Phi_j^n) = A + 2B + T_j^{n+1/2} \geqslant 0. \tag{6.78}$$

对边界 $J_{\partial\Omega}$ 上的点，我们有 $e_j^n = 0$, $\Phi_0^n = \Phi_J^n = At_n$ 和 $\Phi_j^0 = Bx_j(1-x_j)$, 所以在边界上 $e_j^n - \Phi_j^n \leqslant 0$. 因此由引理 6.1, 在 J_Ω 中 $e_j^n - \Phi_j^n \leqslant 0$. 我们现在考虑常数 A 和 B 的两种选取办法.

(i) 令 $A=T$, $B=0$; 因此我们就证明了

$$e_j^n \leqslant \Phi_j^n = t_nT. \tag{6.79}$$

对网格函数 $(-e_j^n - \Phi_j^n)$ 进行相同的论证，就得到

$$|e_j^n| \leqslant t_nT \tag{6.80}$$

这与 (2.96) 吻合.

(ii) 令 $A=0$, $B=\frac{1}{2}T$, 用相同的方法我们得到

$$|e_j^n| \leqslant \frac{1}{2}x_j(1-x_j)T. \tag{6.81}$$

综合这些结果，显然有

$$|e_j^n| \leqslant \max_{\substack{m<n\\0<i<J}}\{|T_i^{m+1/2}|\}\min\left\{t_n, \frac{1}{2}x_j(1-x_j)\right\} \tag{6.82}$$

这也正反映了在区域的边界上 $e=0$, 而随着离开边界误差是逐渐增长的.

我们发现，对 2.6 节的模型问题，随着 t 增长，其解的误差趋于 0, 但是 (6.82) 中的误差界并非如此. 如果我们获知解的更多有关信息，并且我们可以证明 $|T_j^{n+1/2}| \leqslant \tau_n$, 这里 τ_n 是一个已知的关于 n 的减函数，那么我们就有可能构造一个能够导出随 n 递减的误差界的比较函数. 习题 7 中给出了一个例子.

用类似的方法，我们可以得到 2.13 节中赋诺伊曼条件的问题的误差界. 在 $x=0$, 即 x_0 处赋齐次诺伊曼条件，算子在 $j=1$ 处被替换为

$$(L_h\psi)_1^n = \frac{\theta(\psi_2^{n+1} - \psi_1^{n+1}) + (1-\theta)(\psi_2^n - \psi_1^n)}{(\Delta x)^2} - \frac{\psi_1^{n+1} - \psi_1^n}{\Delta t}. \tag{6.83}$$

上式可由 (2.103) 导出，其中令 $\alpha=0$ 及 $g=0$. 由 (2.109), 取 $\theta=0$, 得该点处的截断误差为

$$T_1^{n+1/2}=\frac{1}{2}\Delta t\, u_{tt}-\frac{1}{12}(\Delta x)^2 u_{xxxx}-\frac{1}{2}u_{xx}. \tag{6.84}$$

现在我们运用定理 6.2 的证明方法，并令 J_1 包含所有 $j>1$ 的内部网格点， J_2 包含所有 $j=1$ 的内部网格点. 在 (6.61) 中，我们可以如前取 $T_1=T=\frac{1}{2}\Delta t M_{tt}+\frac{1}{12}(\Delta x)^2 M_{xxxx}$, 但在 J_2, 我们必须采用 $T_2=T+\frac{1}{2}M_{xx}$. 我们构造一个比较函数使得 $\Phi_0^n=\Phi_1^n$, 即取

$$\Phi_j^n=\begin{cases} At_n+B(1-x_j)(1-\Delta x+x_j), & j\in J_1,\\ At_n+B(1-x_j)(1-\Delta x+x_j)+K, & j\in J_2,\end{cases} \tag{6.85}$$

于是我们得到

$$L_h\Phi_j^n=\begin{cases} -(A+2B), & j\in J_1,\\ -(A+2B)-K/(\Delta x)^2, & j\in J_2.\end{cases} \tag{6.86}$$

这证明，如果我们选取常数使

$$\begin{aligned} A+2B&\geqslant T,\\ A+2B+K/(\Delta x)^2&\geqslant T+\frac{1}{2}M_{xx}, \end{aligned} \tag{6.87}$$

成立，则 $L_h(e_j^n-\Phi_j^n)\geqslant 0$ 在 J_1 和 J_2 上均成立. 显然，如果我们取与以前相同的 A 和 B, 以及 $K=\frac{1}{2}(\Delta x)^2M_{xx}$, 要求即得到满足.

区域的边界只包含了 $j=J$ 或 $n=0$ 的点，并且在这些点处，同以前一样，容易看出 $e_j^n-\Phi_j^n\leqslant 0$, 因此它也在内点成立. 于是

$$e_j^n\leqslant At_n+B(1-x_j)(1-\Delta x+x_j)+\frac{1}{2}(\Delta x)^2M_{xx}. \tag{6.88}$$

选取与以前相同的 A 和 B, 我们就得到最终的结果

$$|e_j^n|\leqslant \max_{\substack{m<n\\0<i<J}}\left\{\left|T_i^{m+1/2}\right|\right\}\times\min\left\{t_n,(1-x_j)(1-\Delta x+x_j)+\frac{1}{2}(\Delta x)^2M_{xx}\right\}. \tag{6.89}$$

这说明误差为 $O(\Delta t)+O((\Delta x)^2)$, 恰如赋狄利克雷边界条件的情形.

6.6　渐近误差分析

在上一节中，我们给出了这样的例子：直接的误差分析给出的误差界关于网格尺寸是一阶的，而更精巧的分析给出了一个二阶的误差界. 这必然导致这样的疑问，更细致的误差分析是否能证明误差事实上是三阶的. 在那些例子中，很显然，误差阶不可能再改进了，但对更复杂的问题，看出误差实际的阶到底是多少远非易事. 比方说，在前面的例子中，困难来自于边界附近特殊点处的低阶截断误差，但在考虑 2.15 节， 3.5 节和

下一节中更一般的椭圆型算子时，在每个点处都会有低阶截断误差．因此，当网格尺寸趋于零时，那些能够更精确地刻画误差在极限情形的性态的估计是有用的．

这种估计常常会更充分地利用引理 6.1 的最大值原理，以及对椭圆型算子 L 成立的相应结果．作为预备，假设用 $\Phi_{\partial\Omega}$ 来表示出现在 (6.52) 的误差界中的 Φ_A 在所有边界节点上的最大值，其中 Φ 满足 (6.51). 现在我们对 $\Psi := e - T(\Phi_{\partial\Omega} - \Phi)$ 应用引理 6.1, 显然有

$$L_h\Psi_P = -T_P + TL_h\Phi_P \geqslant 0,$$

并且 Ψ 在边界上的最大值为零，所以我们可以得到结论 $\Psi_P \leqslant 0, \forall P \in J$. 我们还可以对 $-e$ 重复此论证，并且由此推出

$$|e_P| \leqslant T(\Phi_{\partial\Omega} - \Phi_P), \tag{6.90}$$

这个结果比定理 6.1 的 (6.52) 强不少．特别地，它给出了一个在边界上某点递减到零的误差界．

为阐明如何估计误差的渐近性态，我们首先考虑单位正方形上边界赋狄利克雷条件的泊松方程的解．采用标准的五点差分格式，不难写出其截断误差的表达式，并且由于在适当的边界条件下，原问题的精确解是光滑的，所以我们在展开式中多取几项，

$$T_{r,s} = \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy})_{r,s} + \frac{1}{360}(\Delta x)^4(u_{xxxxxx} + u_{yyyyyy})_{r,s} + \cdots. \tag{6.91}$$

于是误差 $e_{r,s}$ 满足方程

$$\begin{aligned} L_h e_{r,s} &= -T_{r,s} \\ &= -\frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy})_{r,s} + O((\Delta x)^4). \end{aligned} \tag{6.92}$$

我们现在假设 $\psi(x,y)$ 为方程

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}) \tag{6.93}$$

的解，它在单位正方形边界上取值为零；并且令 Ψ 为用我们的数值格式近似这个问题所得到的结果，则应用定理 6.1 的误差界可证明 $\Psi - \psi = O((\Delta x)^2)$. 进一步，我们把这个结果与 (6.92) 结合起来，再次应用该定理得

$$\frac{e_{r,s}}{(\Delta x)^2} = \psi(x_r, y_s) + O((\Delta x)^2). \tag{6.94}$$

即原泊松方程的数值解的误差由下式给出：

$$U_{r,s} = u(x_r, y_s) + (\Delta x)^2\psi(x_r, y_s) + O((\Delta x)^4). \tag{6.95}$$

这说明误差恰为二阶，不可能更高了，除非在函数 ψ 恒为零的非常特殊的情况下．当然，该表达式仅在网格尺寸趋于零的极限过程中有意义．

我们可以在两个具有不同网格尺寸的网格上求解差分方程，并通过比较两个数值解在共有网格点上的值估计误差的大小．另一种做法是，我们可以用近似解 U 的差商估计

(6.93) 的右端项，然后通过再次实际求解离散方程得到我们此前称为 ψ 的近似解；将其代入 (6.95) 便得到四阶逼近．这个步骤称为 延迟校正(deferred correction)．我们要再次强调，只有网格尺寸足够小使得 (6.95) 的渐近展开有效时，额外的精度阶才可能得到．

上一节的最后一个例子中，在正方形的一条边上赋了诺伊曼边界条件，通过相当任意地选择比较函数 Φ, 即可以得到其误差界，该问题的渐近分析可用相似的方法得出．显然，误差界 (6.75) 几乎不可能是最优的．在这种情况下，误差满足 $L_h(e_{r,s}) = -T_{r,s}$, 这里 $T_{r,s}$ 由 (6.68) 给出．如果现在我们定义 ψ 为问题

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}) \tag{6.96}$$

且

$$\psi(0,y) = \psi(x,0) = \psi(x,1) = 0,\ \psi_x(1,y) = -\frac{1}{6}u_{xxx}(1,y), \tag{6.97}$$

的解，我们发现对这个问题应用相同的数值方法，在若干额外的高阶截断误差项的意义下， $e_{r,s}$ 满足由此得到的方程．细节留作练习 (见习题 6)．于是，如同前面的例子，我们得到

$$U_{r,s} = u(x_r, y_s) + (\Delta x)^2\psi(x_r, y_s) + o((\Delta x)^2), \tag{6.98}$$

并且 $\psi(\cdot,\cdot)$ 可用延迟校正方法估计．

以上方法可以直接推广到含有曲线边界的问题，与定理 6.2 应用的第一个例子相同，我们把网格点集分成 J_1 和 J_2, 其中 J_2 由那些有一个或多个邻点位于边界上的网格点组成．在 J_1 中的点处，我们有，如同 (6.92),

$$\left|L_h(e_{r,s}) + \frac{1}{12}(\Delta x)^2(u_{xxxx} + u_{yyyy})\right| \leqslant K_3(\Delta x)^4, \qquad \text{在 } J_1 \text{ 中}, \tag{6.99}$$

其中 $K_3 = \frac{1}{360}(M_{xxxx} + M_{yyyy})$. 现在我们假设曲线边界上给定了狄利克雷边界条件，那么在 J_2 中的点处，我们可以利用 (6.56) 和 (6.57b) 给出

$$|L_h(e_{r,s})| \leqslant K_2\Delta x, \qquad \text{在 } J_2 \text{ 中}. \tag{6.100}$$

如同 (6.93), 我们定义 $\psi(x,y)$ 为方程

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}) \tag{6.101}$$

的解，其在边界上的值为零．则只要函数 u 和 ψ 具有足够多的有界的各阶导数，就存在常数 K_4 和 K_5, 使得

$$\left|L_h(\psi_{r,s}) + \frac{1}{12}(u_{xxxx} + u_{yyyy})\right| \leqslant K_4(\Delta x)^2, \qquad \text{在 } J_1 \text{ 中}.$$

$$|L_h(\psi_{r,s})| \leqslant K_5\Delta x, \qquad \text{在 } J_2 \text{ 中}. \tag{6.102}$$

从 (6.99) 和 (6.100) 中减去 (6.102), 我们得到

$$\left|L_h\left(\frac{e_{r,s}}{(\Delta x)^2}-\psi_{r,s}\right)\right|\leqslant (K_3+K_4)(\Delta x)^2, \qquad \text{在 } J_1 \text{ 中},$$

$$\left|L_h\left(\frac{e_{r,s}}{(\Delta x)^2}-\psi_{r,s}\right)\right|\leqslant \frac{K_2}{\Delta x}+K_5\Delta x, \qquad \text{在 } J_2 \text{ 中}. \tag{6.103}$$

我们现在应用定理 6.2; 函数 Φ_P 以及 C_1 和 C_2 的值与那些用以得到 (6.64) 的相同, 即 $C_1=4E_1$, $C_2=E_2/(\Delta x)^2$, 但是 $T_1=(K_3+K_4)(\Delta x)^2$, $T_2=K_2/\Delta x+K_5\Delta x$. 这样我们就得到

$$\left|\frac{e_{r,s}}{(\Delta x)^2}-\psi_{r,s}\right|\leqslant (E_1R^2+E_2)\max\left\{\frac{(K_3+K_4)(\Delta x)^2}{4E_1},\frac{K_2+K_5(\Delta x)^2}{E_2/\Delta x}\right\}. \tag{6.104}$$

选择 $E_1=\frac{1}{4}(K_3+K_4)(\Delta x)^2$, $E_2=K_2\Delta x+K_5(\Delta x)^3$, 得

$$\left|\frac{e_{r,s}}{(\Delta x)^2}-\psi_{r,s}\right|\leqslant C(\Delta x), \tag{6.105}$$

这表明

$$e_{r,s}=(\Delta x)^2\psi(x_r,y_s)+O((\Delta x)^3). \tag{6.106}$$

这也再一次说明在边界附近的点的低阶截断误差是如何不影响渐近展开式的误差主项的.

综合本节得到的渐近误差估计和前几节的严格误差界, 就可以对我们一直在讨论的具有光滑解的线性问题的误差性态给出有用的描述. 不过, 我们应该注意, 对于含有变系数 $a(x,y)$ 的更一般的方程, 构造具有所要求性质的函数 Φ 可能会更困难. 到底有多困难, 误差界的估计有多精确, 都会与 $a(x,y)$ 的形式有关, 在构造的过程中, 很可能会用到它的下界.

然而, 将上面给出的分析方法应用到在曲线边界的一部分上赋诺伊曼条件的问题的尝试, 并不是那么成功. 如果我们采用 (6.41) 的近似, 并且做泰勒级数展开, 我们发现截断误差的主项是

$$\frac{p_1^2-p_2^2}{1+p_1+p_2}u_{xy}. \tag{6.107}$$

我们立刻发现它的阶是 $O(1)$, 当网格尺寸趋于零时它并不收敛到零. 不难应用最大值原理得到误差的一个界, 我们发现这个界是 $O(\Delta x)$ 的. 但是, 我们的误差渐近展开是建立在这样的事实之上的, 即截断误差的主项是 Δx 的幂与只依赖于 x, y 而与 Δx 无关的函数的乘积. 对表达式 (6.107) 来说, 这已经不成立了, 因为 p_1 和 p_2 都是关于 x, y 和 Δx 的函数, 而且还不是光滑函数, 其中含有如分式 $x/\Delta x$ 的项. 图 6-4 反映了这一观察结果. 这幅图绘出了在半径分别为 1 和 0.3 的圆之间的圆环上泊松方程数值解误差的有关

结果. 在第一个问题中, 两条边界上都赋狄利克雷边界条件, 并且边界条件和 f 的选取使解恰好为

$$u(x,y) = 1 + 3x^2 + 5y^2 + 7x^2y^2 + (x^2+2)^2.$$

在这个计算中最大误差如下方的曲线所示, 表明误差约按 $(\Delta x)^2$ 变化. 在第二个问题中, 解同前, 只是在圆环的外边界上加诺伊曼条件. 我们注意到误差大了许多, 其变化趋势为 $O(\Delta x)$, 但是其细微处的性态非常不规则. 对这样一个问题, 为求得误差性态光滑且大致为 $O((\Delta x)^2)$ 的数值解, 就需要对边界条件做更复杂的近似.

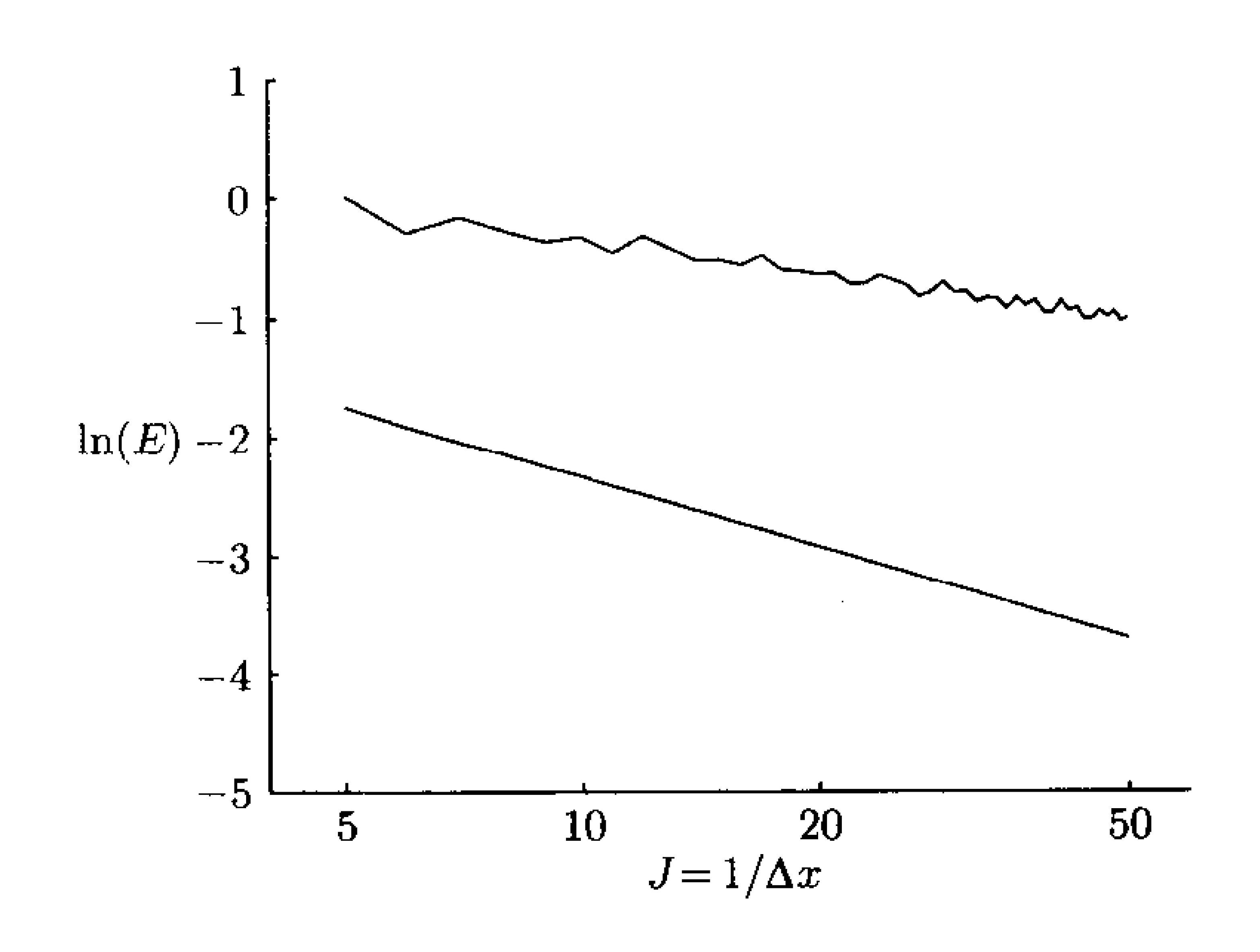


图 6-4　圆环上泊松方程的数值解: 下方曲线, 狄利克雷条件; 上方曲线, 外边界加诺伊曼条件

我们在下一节将看到, 有限元方法的一个主要的优点就是, 它处理诺伊曼边界条件的方式更加简单且自然, 这使误差的性态更佳.

6.7　变分形式和有限元方法

对一般的扩散问题 (6.20) 和 (6.22), 可以给出其变分公式. 首先考虑在边界上所有点处都给出狄利克雷边界条件的情形. 我们定义

$$I(v) := \int_\Omega \left[\frac{1}{2}a|\nabla v|^2 - fv\right] \mathrm{d}x\mathrm{d}y. \tag{6.108}$$

我们断言 (6.20) 的解满足我们写作下面形式的变分方程:

$$\delta I(u) = 0. \tag{6.109}$$

其含义为如果 $v = u + \delta u$ 是任何使 $I(v)$ 有定义, 且满足边界条件的函数, 则

$$I(u+\delta u) - I(u) = o(\delta u). \tag{6.110}$$

我们可以不给出完全严格的证明, 而说明为什么会是这样. 展开 $I(u+\delta u)$, 我们发现

$$
\begin{aligned}
I(u+\delta u)-I(u) &= \int_\Omega [(a\nabla u)\cdot(\nabla\delta u)-f\delta u]\mathrm{d}x\mathrm{d}y+\int_\Omega \frac{1}{2}a|\nabla\delta u|^2\mathrm{d}x\mathrm{d}y \\
&= \int_\Omega -[\nabla\cdot(a\nabla u)+f]\delta u\,\mathrm{d}x\mathrm{d}y+O((\delta u)^2), \qquad (6.111)
\end{aligned}
$$

这里我们用到了高斯定理，以及在 Ω 边界上 $\delta u=0$ 的事实．由此即可得到所需的结论．

事实上，我们证明的远不止于此，因为既然 $a(x,y)>0$，那么对于所有在边界上为零的函数 δu，由 (6.111) 知 $I(u+\delta u)\geqslant I(u)$．因此，当 $I(v)$ 取遍所有满足边界条件的 v 时，函数 u 使其取到了最小值．

下面假设我们取以下形式的有限项展开式：

$$
V(x,y)=\sum_{j=1}^{N}V_j\phi_j(x,y), \qquad (6.112)
$$

其中函数 ϕ_j 是给定的，然后我们试图选择系数 V_j，以使得上式给出解 u 的一个好的近似．我们可以把 $I(V)$ 展开成以下形式

$$
I(V)=\frac{1}{2}\sum_i\sum_j A_{ij}V_iV_j-\sum_i b_iV_i, \qquad (6.113)
$$

其中

$$
A_{ij}=\int_\Omega [a\nabla\phi_i\cdot\nabla\phi_j]\mathrm{d}x\mathrm{d}y \qquad (6.114)
$$

以及

$$
b_i=\int_\Omega f\phi_i\,\mathrm{d}x\mathrm{d}y. \qquad (6.115)
$$

由于精确解 u 使 $I(v)$ 取到最小值，因此通过选择系数 $\{V_j\}$，使 $I(V)$ 取到最小值来定义解的近似 U 是自然的做法．如 (6.113) 所示，$I(V)$ 是系数的二次型，在 V 满足边界条件的约束下，确定其最小值是一件直截了当的事情．

Rayleigh, Ritz 和其他人在十九世纪都用过这种方法，他们选择了各种各样的函数 $\phi_j(x,y)$．这也是现在广泛用来求解椭圆型问题的，比有限差分方法更受人们 (特别是工程师们) 喜爱的有限元方法的起点．

有限元方法独特的特征在于其对 $\phi_j(x,y)$ 的选择，它们通称为*试探函数*(trial function) 或*形函数*(shape function)，其选择应使每个 ϕ_j 仅在区域 Ω 的一小部分上非零．于是在矩阵 A 中，大部分元素都将为零，因为仅当形函数 ϕ_i 与 ϕ_j 相重叠时 A_{ij} 才是非零的．下面我们将要考虑一种简单情况．

假设区域 Ω 是一个多边形，并且剖分为三角形单元，如图 6-5 所示．三角形的顶点 P 通称为 *节点* (node)．假设 V 是分段线性的，其在每个三角形上的线性形式由它在三角形的顶点处的值决定．则显然，它在相邻的三角形间是连续的．我们假设沿着多边形的边界，边界条件类似地也是分段线性的，而 V 就取这些值．函数 $\phi_j(x,y)$ 就定义为在节点 P_j 处取 1, 而在其他节点处取 0 的分段线性函数．这样定义的函数是唯一的，而且仅在以 P_j 为一个顶点的三角形上不为零．此函数被称为 *帽形函数*(hat function), 其原因是不言而喻的． 画在三维空间中 (图 6-6), 它具有金字塔般的形状，每个面都是三角形．

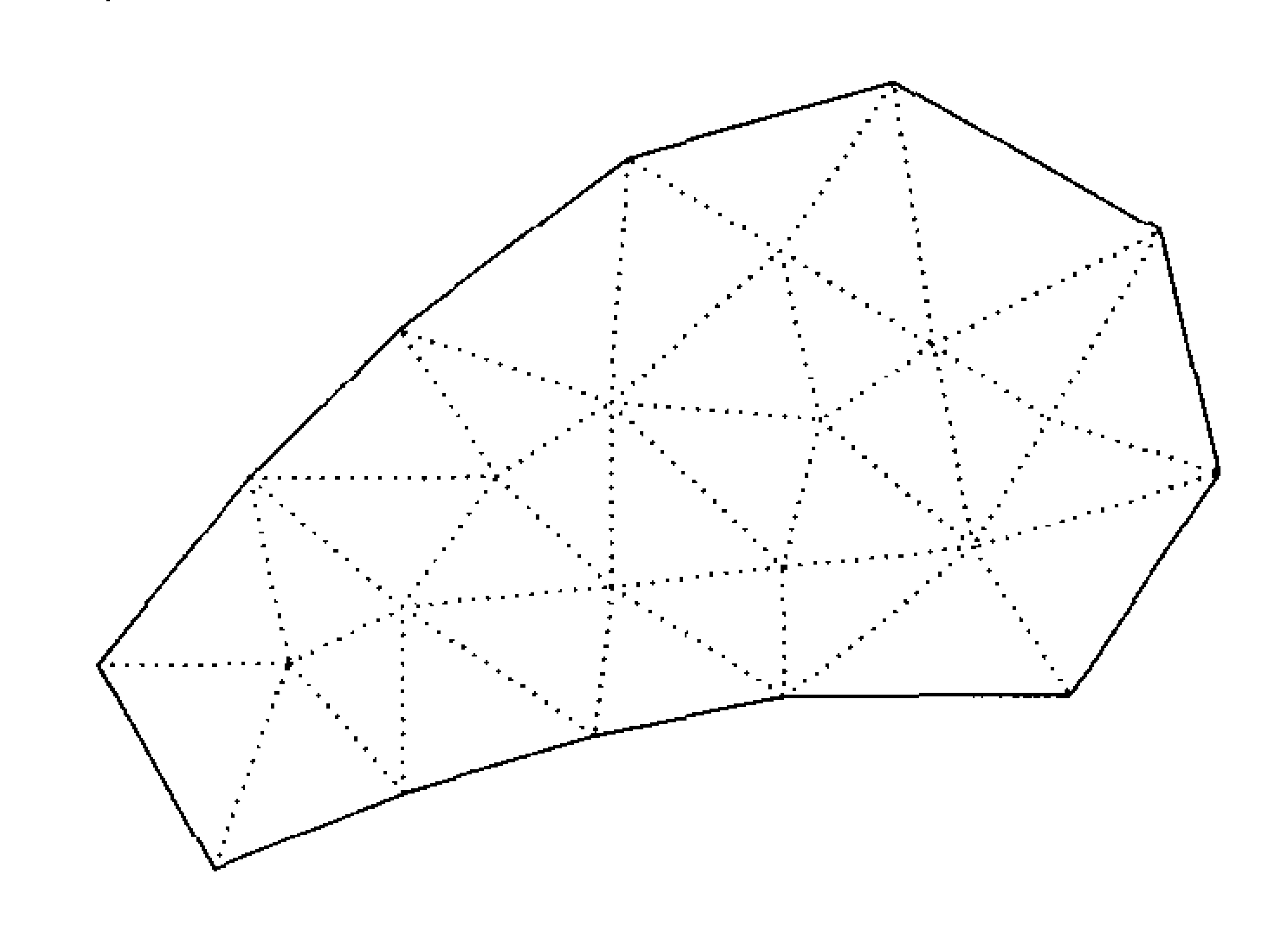


图 6-5 多边形上的三角形单元

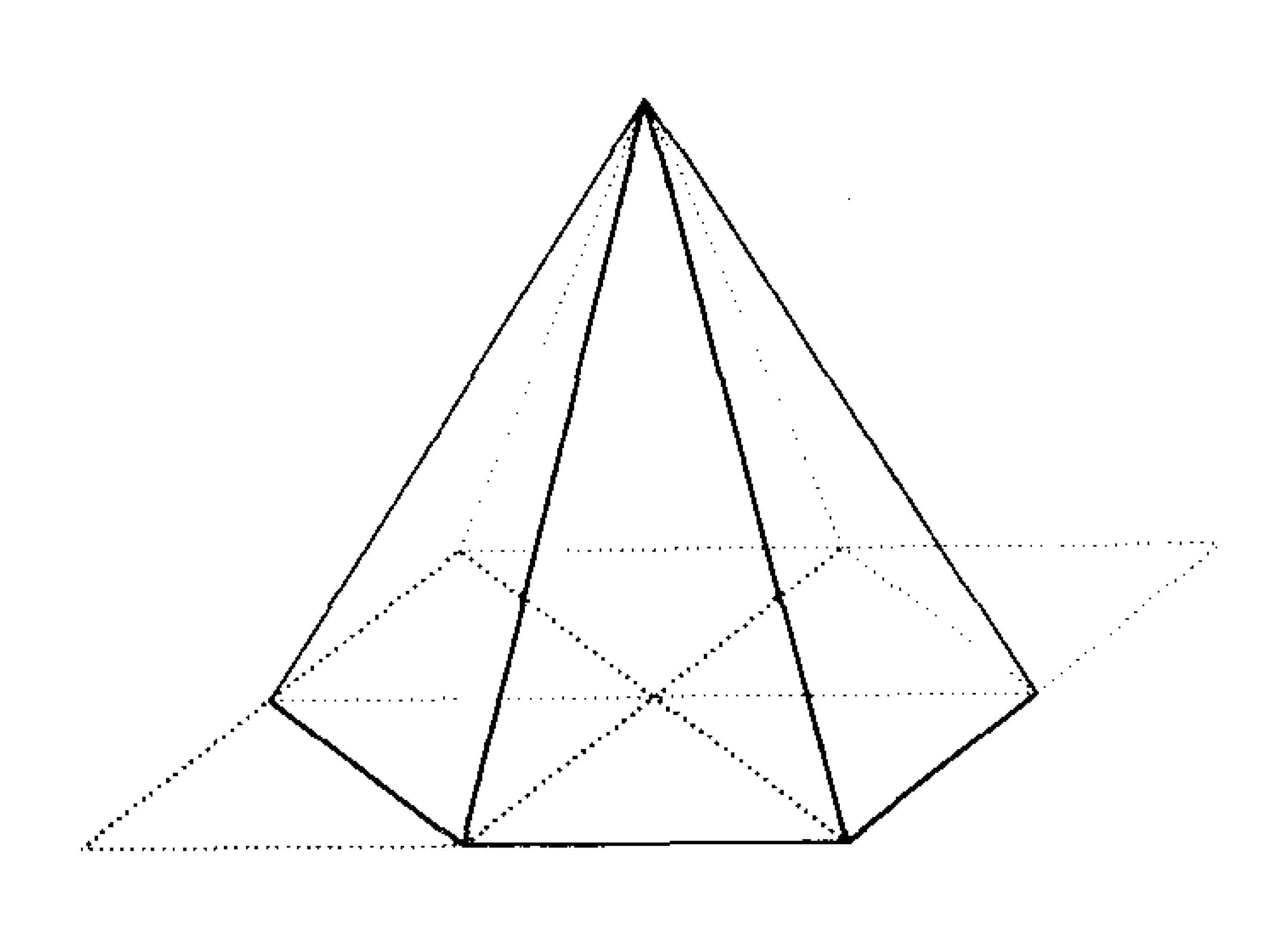


图 6-6 一个帽形基函数

这样定义的形函数具有一个有用的性质，就是系数 V_j 给出了函数 $V(x,y)$ 在节点 P_j 处的值，这是因为所有其他的形函数在这个节点处的值都为零．为了确保 V 满足边界条件，我们固定对应于边界上节点的系数 V_j, 而允许其他系数在求最小值的过程中变化．

由于三角形剖分可以容易地拟和几乎任何形状的边界，有限元方法的一个主要的优点就是它可以相当容易地适用于复杂的几何形状．一旦构造了一组三角形单元，矩阵 A 的元素就可以通过一套程式化的步骤计算，而不用考虑三角形的具体形状．

为说明这套步骤，我们来考虑最简单的模型问题，如 (6.1), 在一个正方形上求解边界

上 $u=0$ 的泊松方程. 如图 6-7 所示， 我们先用以前的做法，在此正方形上构造一个一

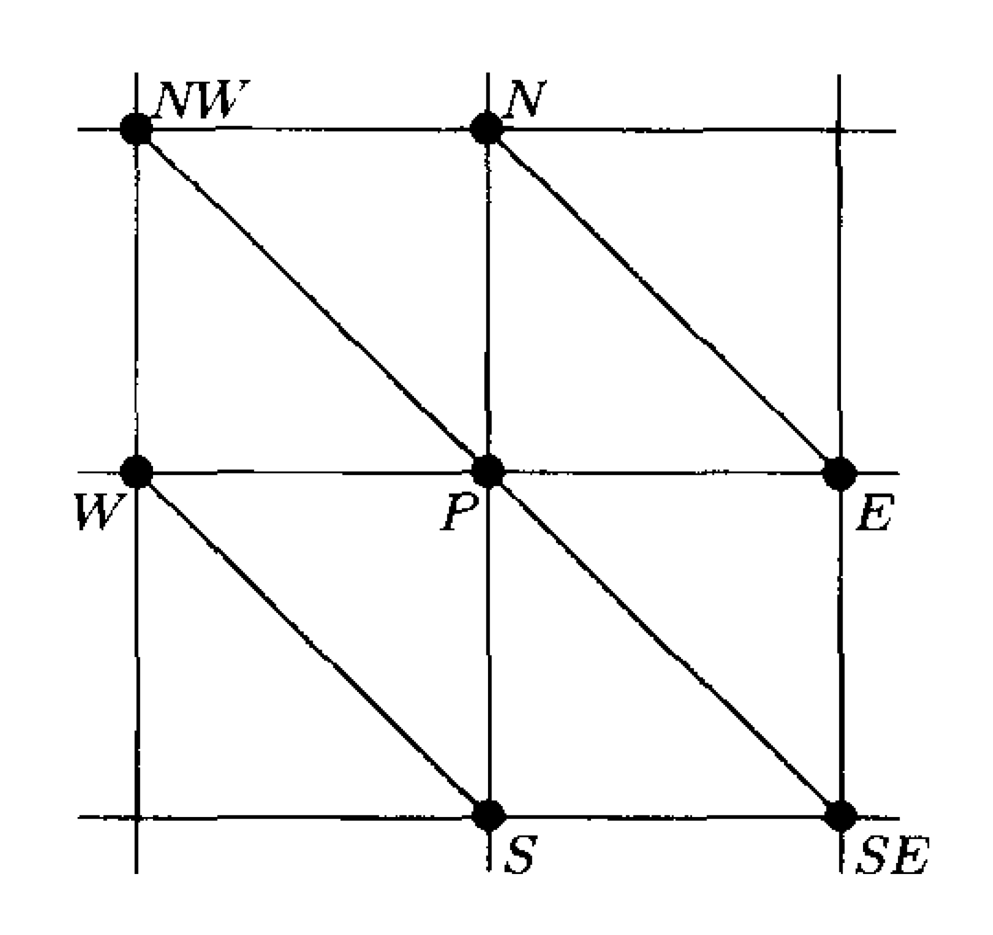


图 6-7 正方形网格上的三角形单元

致的正方形网格，然后在每个小正方形上画出其对角线，将其分为两个三角形. 于是在矩阵 A 中，每行一般包含 7 个非零元素，因为中心在 P 的帽形函数会与中心在 P 以及 P 和它的六个相邻的节点的帽形函数交叠，这六个节点在图上标记为 N, S, E, W, NW 和 SE 的. 由于每个 ϕ 在各三角形上都是关于 x, y 的线性函数，其梯度 $\nabla\phi$ 在每个三角形上都是常数. 在这个例子中 $a(x,y)=1$, 因此计算

$$A_{ij}=\int_{\Omega}[\nabla\phi_i\cdot\nabla\phi_j]\mathrm{d}x\mathrm{d}y \tag{6.116}$$

是一件很简单的事. 在每个三角形上， ϕ 的各偏导数取值均为 $0, 1/\Delta x$ 或 $-1/\Delta x$ 之一，且有

$$A_{PP}=4, \tag{6.117a}$$

$$A_{PN}=A_{PS}=A_{PE}=A_{PW}=-1, \tag{6.117b}$$

及

$$A_{P,NW}=A_{P,SE}=0. \tag{6.117c}$$

我们还需要由 (6.115) 计算 b_i 的值. 如果我们将函数 $f(x,y)$ 用 f 在 P 的值这个常数代替，则得到 b_P 近似为

$$b_P=(\Delta x)^2 f_P. \tag{6.118}$$

不过，我们要在这里指出，在有限元程序中，积分 (6.114) 和 (6.115) 首先在每个单元上完成，然后再组装成一个具有下面 (6.119) 形式的整体方程，特别地，这就意味着，在每个单元上， (6.115) 通常将用重心积分公式近似，所以 (6.118) 中的 f_P 将替换为以 P 为一顶点的六个三角形重心处的值的平均. 二次型 (6.113) 的最小值由满足线性方程组 $A\mathbf{U}=\mathbf{b}$ 的向量 $\mathbf{U}$ 给出，对应于可变的 V_j 的内部节点的数目决定了 $\mathbf{U}, \mathbf{b}$ 的，进而 A 的维数. 因

为边界值都为零，所以方程右端项再无其他贡献．该方程组中一个一般的方程为

$$4U_P - U_N - U_S - U_E - U_W - (\Delta x)^2 f_P = 0, \tag{6.119}$$

这与 6.1 节引进的我们熟悉的有限差分格式是一样的，除了整体改变了正负号．

与有限差分方法相比，有限元方法的误差分析有一个不同的特点．我们已经看到，对任何在边界上取值为零的函数 w, u 满足方程

$$\int_\Omega [a\nabla u \cdot \nabla w - fw]\mathrm{d}x\mathrm{d}y = 0. \tag{6.120}$$

而且如果积分是精确的 (或者 a 和 f 是常数), 则对任何可以表示为形如 (6.112) 的有限和并且在边界上取值为零的函数 W, 函数 U 也满足方程

$$\int_\Omega [a\nabla U \cdot \nabla W - fW]\mathrm{d}x\mathrm{d}y = 0. \tag{6.121}$$

现在我们可以取任何一个满足边界条件且具有 (6.112) 形式的函数 V, 并将 w 和 W 均取为 $V-U$, 因为这个差在边界上取值为零．于是，上两式相减，我们便得到

$$\int_\Omega [a\nabla(U-u) \cdot \nabla(V-U)]\mathrm{d}x\mathrm{d}y = 0. \tag{6.122}$$

因此

$$\begin{aligned}\int_\Omega a|\nabla(V-u)|^2\mathrm{d}x\mathrm{d}y &= \int_\Omega a|\nabla[(V-U)+(U-u)]|^2\mathrm{d}x\mathrm{d}y \\ &= \int_\Omega a|\nabla(V-U)|^2\mathrm{d}x\mathrm{d}y + \int_\Omega a|\nabla(U-u)|^2\mathrm{d}x\mathrm{d}y,\end{aligned}$$

这是因为由 (6.122) 含交叉积的项为零．这就意味着

$$\int_\Omega a|\nabla(U-u)|^2\mathrm{d}x\mathrm{d}y \leqslant \int_\Omega a|\nabla(V-u)|^2\mathrm{d}x\mathrm{d}y, \quad \forall V. \tag{6.123}$$

对在 Ω 边界上取值为零的函数 $w(x,y)$, 我们定义一特殊的范数，

$$\|w\|_E^2 = \int_\Omega a|\nabla w|^2\mathrm{d}x\mathrm{d}y. \tag{6.124}$$

这样我们就已经证明了

$$\|U-u\|_E \leqslant \|V-u\|_E, \quad \forall V. \tag{6.125}$$

这个关键的结果说明 u 具有 (6.112) 形式的所有可能的近似中，在这个范数的意义下， U 是最好的．其误差可通过应用逼近论非常精确地估计出来．在这里涉及任何细节，或考虑 a 和 f 为变量时采用数值积分格式的影响，都会超出本书的范围，因此我们只引用主要的结果如下

$$\|U-u\|_E \leqslant C_1 h|u|_*, \tag{6.126a}$$

$$\|U-u\|_{L_2(\Omega)} \leqslant C_2 h^2|u|_*, \tag{6.126b}$$

其中 h 是三角形剖分的所有三角形的最大直径. $|u|_* = \|u_{xx}\| + \|u_{xy}\| + \|u_{yy}\|$, 而 $\|\cdot\|$ 表示 $\|\cdot\|_{L_2(\Omega)}$ 范数.

最后，假设在边界的一部分上加了齐次诺伊曼边界条件. 这种边界条件是以一种自然的形式体现在变分过程的. 而有限元法也同样用这种自然的方式处理这种边界条件. 在导出 (6.111) 时，应用高斯定理得到了边界积分

$$\int_{\partial\Omega} a\frac{\partial u}{\partial n}\delta u\,\mathrm{d}l. \tag{6.127}$$

在所有边界点上，无论是加狄利克雷边界条件，这时 $\delta u = 0$, 还是加齐次诺伊曼边界条件，这种情况下对真解有 $\partial u/\partial n = 0$, 该积分都为零. 因此，当我们求由 (6.108) 式定义的 $I(v)$ 的极小时，在达到最小值处，不仅方程 (6.20) 在所有内点都成立，而且在所有无狄利克雷边界条件约束的边界点处，也有 $\partial u/\partial n = 0$. 这就称作诺伊曼边界条件的 自然(natural) 处理.

有限元方法直接应用了这一求最小值的性质. 不受狄利克雷边界条件约束的边界节点如内部节点一样处理，并且包括在求和式 (6.112) 中，相应的形函数被边界截断. 这样方程组 $A\mathbf{U} = \mathbf{b}$ 的计算可以用同以前一样的方式进行，不同的地方是其维数会因额外增加的节点而增长，而且 A 和 $\mathbf{b}$ 相应的分量会具有非标准形式 (参见 6.9 节的例子). 此外，上面的误差分析过程在此仍然适用；唯一需要更改的是将“满足边界条件”和“在边界处为零”替换为“满足狄利克雷边界条件”和“在狄利克雷边界处取值为零”.

6.8 对流扩散问题

我们在前几章已经看到，在扩散算子上附加低阶项会造成困难，也会使许多标准方法效用降低. 我们在这一章讨论过的微分方程 (6.24) 仍然满足最大值原理，这是因为如果在内部有一个最大值点，那么在该点处就有 $u_x = u_y = 0$, 所以这些附加项不会影响用反证法推导出不存在这样的最大值点. (此外，还要注意，即使这些项与扩散系数 a 无关，这也是成立的.) 但是，引理 6.1 中关于离散方程 (6.25) 的相应的论述就不成立了，除非条件 (6.27) 得以满足，但这个条件限制性很强. 幸运地是，在那里讨论的那种情况下，可以通过应用有限体积法导出差分方程来避免这个困难，由此得到的格式是 (6.29a).

有时一阶项并非因展开扩散项而产生，这时困难就变得严重得多. 刚刚在 6.7 节考虑过的变分形式强调了这种特征：由 (6.20) 中的自伴 (self-adjoint) 微分算子，我们得到了 (6.120) 中的双线性型 (bilinear form):

$$\int_{\Omega} a\nabla u\cdot\nabla w\,\mathrm{d}x\mathrm{d}y. \tag{6.128}$$

它关于 u 和 w 是对称的，并且由此导出了 (6.124) 中定义的范数. 假设我们有如下形式的**对流扩散**(convection-diffusion) 方程：

$$-\nabla(\epsilon\nabla u)+\mathbf{V}\cdot\nabla u=f. \tag{6.129}$$

这里 ϵ 是一个 (正的) 扩散系数，与方程 (6.20) 相比多出来的那些项来自于对流 (或平流) 速度 $\mathbf{V}$, 我们引入了不同的记号用来强调这样一个事实，即扩散系数经常是非常小的. 如果速度是不可压缩的，即 $\nabla\cdot\mathbf{V}=0$, 这个方程就等价于

$$\nabla\cdot(\epsilon\nabla u-\mathbf{V}u)+f=0. \tag{6.130}$$

两种形式都可能在实际问题中出现，而且它们对应了 5.7 节中所考虑的问题的一种二维定常形式；并且它们带来了相同的问题. 用**检验函数** (test function) w 乘后一个方程，然后在其区域上积分，再做分部积分，得双线性型：

$$\int_\Omega(\epsilon\nabla u-\mathbf{V}u)\cdot\nabla w\,\mathrm{d}x\mathrm{d}y. \tag{6.131}$$

它不再关于 u 和 w 对称. 其结果是我们不再有一个可以用来衡量逼近精度的自然的范数，而相应于 (6.121) 的方程组也不再给出 u 的最优逼近.

下面我们将简略地考虑应如何推导有限差分方法来重建最大值原理，以及有限元方法如何能够近乎重获最佳逼近的性质.

导出有限差分格式的最好的途径是通过有限体积法，将针对扩散项的 (6.29) 中的公式和 4.7 节中推出的针对对流项的公式组合在一起. 在后一种情况，我们发现最好采用某种形式的迎风格式，这样，如果速度分量 $\mathbf{V}^{(x)}$ 是正的，在 (6.30) 的罗经点记号下，我们用以下的近似取代东、西方向的通量：

$$\begin{aligned}\left(\epsilon\frac{\partial u}{\partial x}-\mathbf{V}^{(x)}u\right)_e &\approx \epsilon_e\frac{U_E-U_P}{\Delta x}-\mathbf{V}_P^{(x)}U_P,\\ \left(\epsilon\frac{\partial u}{\partial x}-\mathbf{V}^{(x)}u\right)_w &\approx \epsilon_w\frac{U_P-U_W}{\Delta x}-\mathbf{V}_W^{(x)}U_W.\end{aligned} \tag{6.132}$$

如果速率 $\mathbf{V}^{(y)}$ 也是正的，我们对南、北方向的通量做类似的替换，经过这些替换后生成的用于替代 (6.30) 的方程满足引理 6.1 的所有条件，并且其近似解满足最大值原理.

我们认为这是构造迎风格式的最简单的做法，因为尽管在方程 (6.130) 中没有关于时间的导数，但如果引进关于时间的导数项，那么 ϵ 的正负号要求它必须出现在方程的右边，这样特征线就是从左下方指向右上方. 用这样一种简单的方式来确保最大值原理成立，需要付出的代价是我们只能得到一阶精度. 对流扩散问题经常会出现陡峭的边界层，而这个格式，虽然能保证一个具有单调性的层仍然保持其单调性，但经常给出一个比正确结果厚很多的边界层. 这个效应，以及其量值的估计易由修正方程分析得出，这种分析还表明，对流项的一阶近似截断误差会增强扩散项.

在 (6.130) 的有限元逼近中，我们也要引入一些迎风的做法．最好的处理办法是修正用到的检验函数．如果在 (6.121) 式中，用来生成检验函数 W 的基函数 ϕ_j 与生成近似解 U 的展开式 (6.112) 中的一样，那么这种方法称为 *Galerkin*方法(Galerkin method)．采用更一般的检验函数的方法称为 *Petrov-Galerkin*方法(Petrov-Galerkin method)．我们首先在一维情形考虑对对流—扩散问题有用的取法．

图 6-6 中勾画的分段线性基函数，在一维情形为

$$\phi_j(x) = \begin{cases} (x - x_{j-1})/h_j, & x \in e_j, \\ (x_{j+1} - x)/h_{j+1}, & x \in e_{j+1}, \end{cases} \tag{6.133}$$

其中具有长度 h_j 的单元 e_j 对应区间 (x_{j-1}, x_j)．现在假设我们引入了迎风检验函数

$$\psi_j(x) = \phi_j(x) + \alpha\sigma_j(x), \tag{6.134}$$

其中修正项 $\sigma_j(\cdot)$ 定义为

$$\sigma_j(x) = \begin{cases} 3(x - x_{j-1})(x_j - x)/h_j^2, & x \in e_j, \\ -3(x - x_j)(x_{j+1} - x)/h_{j+1}^2, & x \in e_{j+1}. \end{cases} \tag{6.135}$$

这样，如果 α 是正的，那么在 x_j 的左侧 $\phi_j(\cdot)$ 加上了一个二次冒泡函数，同时在右侧减去一个类似的冒泡函数．现在将这种检验函数应用在相应于 (6.121) 的方程

$$\int [\epsilon U' - VU]\psi_j' \mathrm{d}x = \int f\psi_j \mathrm{d}x. \tag{6.136}$$

为简单起见，我们假设 ϵ 和 V 为常数．那么，由于 U' 在每个区间上都是常数，并且 $\sigma_j(\cdot)$ 在每个节点上都为零，因此迎风操作对扩散项没有影响，而我们得到了以下格式

$$\begin{aligned} &-\epsilon \left[\frac{U_{j+1} - U_j}{h_{j+1}} - \frac{U_j - U_{j-1}}{h_j}\right] \\ &+ V\left[\frac{1}{2}(1-\alpha)(U_{j+1} - U_j) + \frac{1}{2}(1+\alpha)(U_j - U_{j-1})\right] = \int f\psi_j \mathrm{d}x. \end{aligned} \tag{6.137}$$

显然，当 V 为正时，取 $\alpha > 0$ 就对应了迎风格式．此外，在这种情况下，对于一致的网格分布，如果

$$\frac{1}{2}(1-\alpha)V \leqslant \epsilon/h, \qquad \text{即} \quad \alpha \geqslant 1 - 2\epsilon/Vh, \tag{6.138}$$

则该格式满足最大值原理．同时也要注意在一个一致网格上，我们可以将 (6.137) 写作

$$-\left(\epsilon/h + \frac{1}{2}\alpha V\right)\delta_x^2 U_j + V\Delta_{0x} U_j = \int f\psi_j \mathrm{d}x. \tag{6.139}$$

这解释了迎风操作是如何增大扩散的．正如我们以前看到过的，当网格 Péclet 数 $\frac{Vh}{\epsilon}$ 小于 2 时，我们可以取 $\alpha = 0$, 并对对流项采用中心差分．但是对于任何大些的值，我们可以选择 α, 譬如令其使 (6.138) 中的等号成立，来确保最大值原理成立．

形式广泛的迎风 '冒泡函数' 可以 (并且已经) 用来获得上述结果. 这就为将这些技巧推广至高维情形提供了广泛的可行性, $\mathbf{V}\cdot\nabla\phi_j$ 提供了一个很好的用来指示哪里应该加冒泡函数的指示器, 事实上, 这形成了得到广泛应用的 *流线扩散法* (streamline diffusion method) 的起点. 读者可以参看文献注释来寻找关于这些方法及其他方法的参考资料. 在这些参考资料中, 你也会发现, 存在在不同的范数下选择检验函数的方法, 使由此导出的近似解在某个给定的因子的范围内为最优的.

6.9　一个例子

作为一个结束本章的例子, 我们来考虑这样一个问题, 其几何特征与第 3 章中考虑的一个问题类似 (参看图 3-6). 我们希望能够确定一个由两个长柱体组成的一个导体的电容, 其外柱体的截面为正方形, 内柱体的截面为圆形. 由此导出的是一个二维问题, 需要在一个正方形和一个小一点的同心圆之间的区域上求解一个拉普拉斯方程. 这个问题有 3 条对称线, 因此显然仅需考虑如图 6-8 所示的该区域的八分之一就够了, 而且这样做效率要高得多.

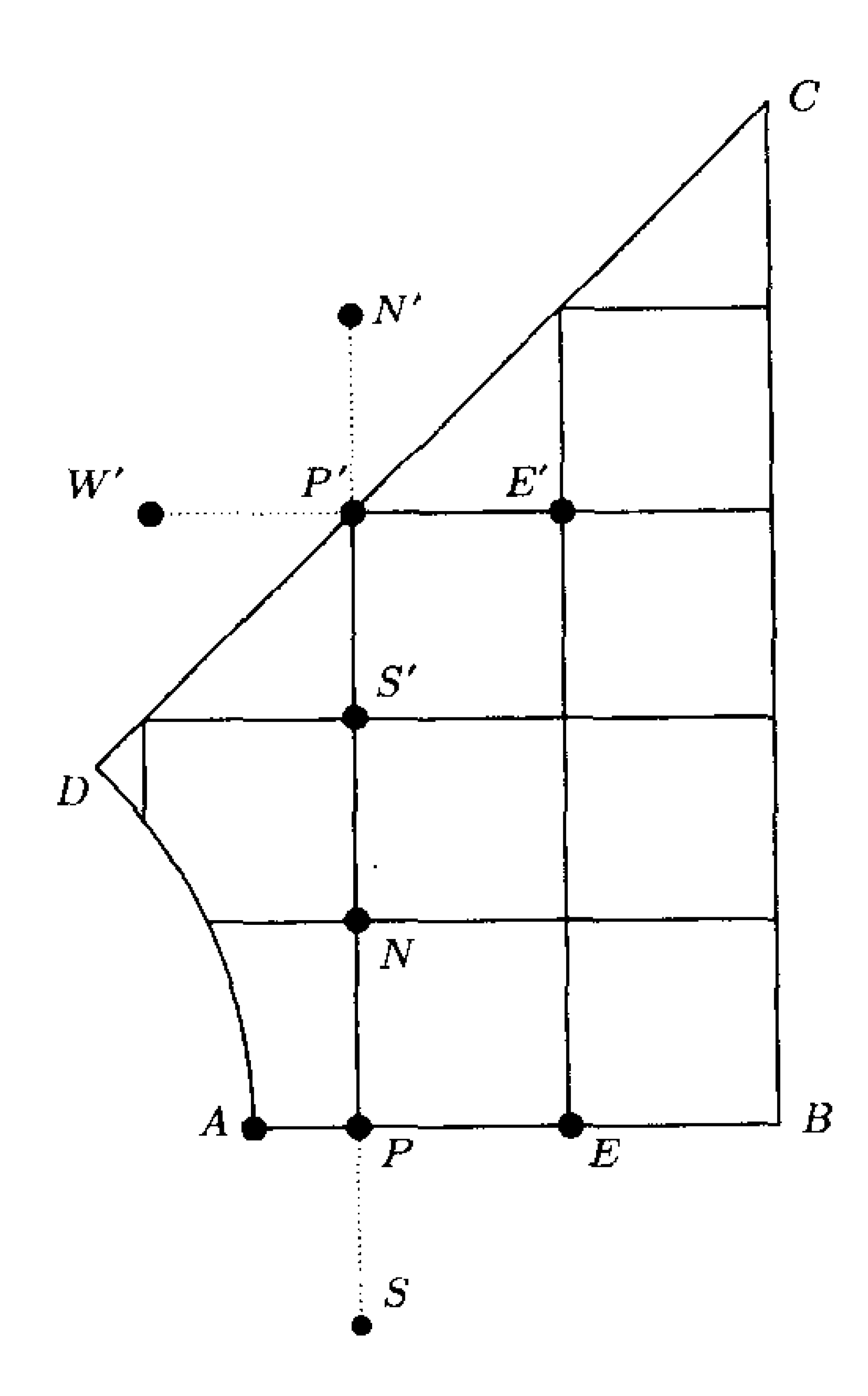


图 6-8　例: 拉普拉斯方程

我们需要在由圆弧 AD 与线段 AB, BC 和 CD 包围的区域上求解拉普拉斯方程. 边界条件规定, 解在圆弧 AD 上为零, 而在线段 BC 上的值为 1. 另外的两条线段是对称边界; 形式上我们可以指定法向导数在这两条线段上为零. 但是有一种更简单并且更精确的直接应用对称性的方法, 就是首先确保这两条边界线经过网格点, 然后将这些网格点

视为内点，并在这些网格点处建立非标准方程.

如果应用有限差分格式，对对称边界 AB 上的一点 P, 我们构造其在边界外的相邻网格点 S. 由对称性，$U_S = U_N$, 所以，相应于点 P 的差分方程成为

$$c_N U_N + c_S U_N + c_E U_E + c_A U_A - c_P U_P = 0. \tag{6.140}$$

类似地，在对称边界 CD 处，我们构造边界外的点 N' 和 W', 这样应用对称性于

$$c_{N'} U_{N'} + c_{S'} U_{S'} + c_{E'} U_{E'} + c_{W'} U_{W'} - c_{P'} U_{P'} = 0 \tag{6.141}$$

得

$$(c_{W'} + c_{S'}) U_{S'} + (c_{E'} + c_{N'}) U_{E'} - c_{P'} U_{P'} = 0. \tag{6.142}$$

在除了与圆弧邻近的节点外的所有其他节点处应用标准差分格式，与圆弧邻近的节点的差分格式的系数可根据 6.4 节讨论过的方法计算.

如果对这个问题应用有限元方法，在边缘 AB 和 CD 内部节点处的值包括在求极小值的过程中. 每个这样的节点都有一个在边界处被截断的相应的形函数，并且由此导出一个修正方程. 假设生成三角形单元的对角线是画在 CD 方向上的，那么对于 CD 上的典型点 P', 由对称性，显然有系数 $c_{P'}$ 减小一半，而 $c_{E'}$ 和 $c_{S'}$ 不因形函数被截断而变化. 所以我们得到的方程是将 (6.142) 折半. 在 AB 上的点 P 处，系数 c_P,c_E 和 c_W 都减半，而 c_N 不变；所以我们又得到了 (6.140) 折半的方程.

因此在这种情况下，有限差分方法和有限元方法给出了相同的方程组. 不过，后者显然给出了一个比 6.4 节所描述的直接得多的处理诺伊曼边值条件的更一般的处理方法. 图 6-9 所示的是一个在单位正方形和半径为 0.4 的圆所围成的区域上的解. 为了便于可视化，将解关于对角线作了镜像反射.

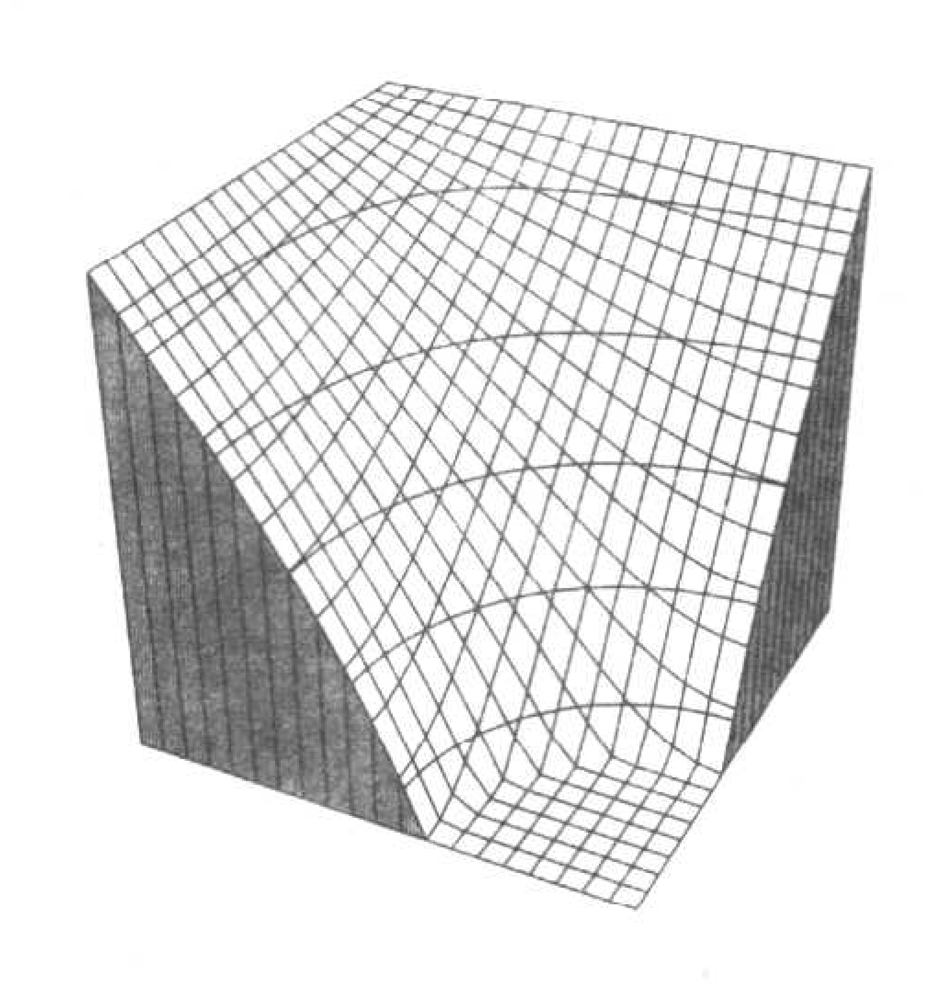


图 6-9 四分之一区域上的解

这个问题需要的结果是边界外电场的通量. 这可由通过圆弧 AD 或者直线段 BC 的通量给出，后者显然更容易精确地计算；由于我们只考虑了区域的八分之一，因此要计算

$$g = 8 \int_{BC} \frac{\partial u}{\partial x} \mathrm{d}y. \tag{6.143}$$

应用中心差分，在点 (x_r, y_s) 处，它由

$$\frac{U_{r+1,s} - U_{r-1,s}}{2\Delta x}$$

近似，其中的一个点在边界之外．在边界点处应用标准的五点差分格式近似拉普拉斯方程，我们可以消去 $U_{r+1,s}$，从而得到近似

$$\frac{4U_{r,s} - 2U_{r-1,s} - U_{r,s+1} - U_{r,s-1}}{2\Delta x}.$$

除了 $\partial u/\partial n$ 假设为零的 C 点，在其余点用此式近似法向导数，通过应用斜方形法则，容易由 (6.143) 计算得到要求的通量．

请读者作为练习，来验证一下按照计算 (6.116) 积分相同的方法计算的如下形式的有限元积分会导出相同的公式．

$$\int_{\Omega} [\nabla U \cdot \nabla W] \mathrm{d}x\mathrm{d}y, \tag{6.144}$$

其中 W 是在 BC 上为单位值，而在穿过 E 的垂线上为零的分段线性函数．

表 6-1 列出了关于各种网格尺寸的数值结果．准确值是 6.33362，它们清晰地显示了一个 $(\Delta x)^2$ 阶的误差．事实上，该数值结果与下面的表达式有四位小数是吻合的

$$6.33362 + 0.53(\Delta x)^2.$$

表 6-1　对不同的 Δx 计算的通量值

Δx	g
0.1000	6.33901
0.0500	6.33495
0.0250	6.33396
0.0125	6.33370

文献注记与推荐读物

前面已经提到过的 Collatz(1966) 的经典教材中，给出了对处理带一般边界条件的椭圆问题的方法的分析．通过在小面积单元上积分来构造差分方程的思想，似乎是由 Varga(1974) 首次进行广泛宣传的．

最大值原理在有限差分方法的误差分析中的应用，是由 Bramble 和 Hubbard 在其始于 1964 年的一系列文章中发展起来的，在参考资料中已经有选择地提及了其中的一部分. 关于最大值原理对椭圆型方程更一般的应用，可参见 Protter 和 Weinberger (1967) 的教材.

我们对有限元方法所做的简略介绍，确实没有足够地反映出其对椭圆型问题求解的影响，而那些更深的理论已涵盖于许多书中，例如，见 Strang 和 Fix (1973), Ciarlet (1978), Brenner 与 Scott (2002) 的论著. 这些书涵盖了许多更深的理论，包括高阶形函数，四边形单元和曲边单元，以及在各种范数下的误差分析.

有限元方法的一个关键的特点是，它使得网格的局部加密或者在选定的单元上引进高阶形函数成为可能. 这种在选定的单元上 (误差比平均误差大) 进行局部加密的方法是基于*后验误差估计* (a posteriori error estimate), 所谓后验误差估计指的是由计算得到的近似解推出的误差估计. Süli 和 Mayers (2003) 中有关于这些思想的简单介绍.

在对流扩散问题中遇到的特殊困难 (包括应用有限体积法和有限元方法) 可以在 Roos *et al.*(1996) 和 Morton(1996) 的书中找到.

习　题

6.1　设泊松方程

$$u_{xx} + u_{yy} + f(x,y) = 0$$

定义在由 $y=0, y=2+2x$ 和 $y=2-2x$ 围成的三角形上，边界上所有点处都给定了狄利克雷边界条件. 为求其数值解，我们采用一致正方形网格，其尺寸为 $\Delta x = \Delta y = 1/N$. 试找出在内点和邻近边界的点处，标准五点差分格式的截断误差的主项. 说明如何选择常数 C, 以便可以对网格函数 $u(x_i,y_j) - U_{i,j} + Cy_j^2$ 应用最大值原理. 证明解的误差关于网格尺寸至少是一阶的.

采用 $\Delta y = 2\Delta x$ 的矩形网格，可以避免在边界附近采用一种特殊的差分格式的必要性，这与前面的方法相比真的具有优越性吗？

6.2　我们要在单位正方形上求解方程

$$\frac{\partial}{\partial x}\left(a\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a\frac{\partial u}{\partial y}\right) + f(x,y) = 0,$$

边界上给定狄利克雷条件. 在一致网格上应用差分格式

$$L_h U := \frac{\delta_x(a\delta_x U)}{(\Delta x)^2} + \frac{\delta_y(a\delta_y U)}{(\Delta y)^2} = -f.$$

假设 a 是只依赖于 x 的正的单调增函数. 说明如何选择常数 C, 使得比较函数 $\Phi = C(x^2+y^2)$ 满足定理 6.1 的条件. 将此结果推广到下面几种情形： (i) a 是只依赖于 x 的正的单调减函数;

(ii) $a(x,y)$ 对固定的 y 是关于 x 的正的单调增函数，并且对固定的 x 是关于 y 的单调减函数.

6.3　直接完整地构造出由 $x \geqslant 0, y \geqslant 0, x^2 + y \leqslant 1$ 定义的区域上，泊松方程

$$u_{xx} + u_{yy} + f(x,y) = 0$$

的近似解的线性方程组. 网格采用尺寸为 $\frac{1}{3}$ 的一致正方形, 边界条件是 $u(x,0) = p(x)$, $u_x(0,y) = q(y)$, 以及 $u(x, 1-x^2) = r(x)$, 其中 p, q, r 和 f 是给定的函数.

6.4　在矩形 $0 \leqslant x \leqslant a, 0 \leqslant y \leqslant b$ 上求解方程

$$u_{xx} + u_{yy} - Ku + f(x,y) = 0,$$

其中 K 是一个正常数，f 是一给定函数；边界所有点都给定狄利克雷边界条件. 我们在尺寸为 $\Delta x = a/(M+1), \Delta y = b/(N+1)$ 的一致网格上应用中心差分近似.

写出截断误差的主项，推演最大值原理. 证明通过合理选择常数 C 和 D, 就可以对网格函数 $U_{r,s} - u(x_r, y_s) + Cx_r^2 + D$ 应用最大值原理以给出数值解的误差界.

6.5　在扇形区域 $0 \leqslant x^2 + y^2 \leqslant 1, 0 \leqslant y \leqslant x$ 上，函数 $u(x,y)$ 满足方程 $u_{xx} + u_{yy} + f(x,y) = 0$. 在边界 $y = x$ 上给定了诺伊曼边界条件，而在其余边界上给定了狄利克雷边界条件. 采用尺寸为 $\Delta x = \Delta y = \frac{1}{3}$ 的一致网格，得到具有形式 $A\mathbf{u} = \mathbf{b}$ 的线性方程组. 明确地写出矩阵 A 的各元素.

将原问题写成极坐标形式，并且构造类似的线性方程组的系数矩阵.

6.6　设 u 是单位正方形上方程 $u_{xx} + u_{yy} = f$ 的解，在 $x = 1$ 的边界上满足诺伊曼边界条件 $u_x = g$, 在另外三条边上满足狄利克雷边界条件. 在一致正方形网格上用通常的中心差分近似，边界条件近似为

$$\frac{U_{J+1,s} - U_{J-1,s}}{2\Delta x} = g_s.$$

证明由此得到的数值解满足

$$U_{r,s} = u(x_r, y_s) + (\Delta x)^2 \psi(x_r, y_s) + O((\Delta x)^3),$$

其中 ψ 满足方程

$$\psi_{xx} + \psi_{yy} = -\frac{1}{12}(u_{xxxx} + u_{yyyy}).$$

试给出 $\psi(x,y)$ 满足的边界条件.

对 $\Delta y = \frac{1}{2}\Delta x$ 的一致矩形网格推导类似结果.

6.7　用 θ- 方法 来求解在 $x = 0$ 和 $x = 1$ 处赋狄利克雷边界条件的热方程 $u_t = u_{xx}$; 假设给定了截断误差的界

$$|T_j^{n+1/2}| \leqslant C(1 - \alpha\Delta t)^n, \quad n \geqslant 0,\ 0 < j < J,$$

其中 α 和 C 是使 $\alpha\Delta t < 1$ 和 $\theta\alpha\Delta t < \frac{1}{2} - \frac{1}{8}\alpha$ 的正常数. 用如下形式的比较函数

$$(1-\alpha\Delta t)^n x_j(1-x_j)$$

导出误差界

$$|e_j^n| \leqslant K(1-\alpha\Delta t)^n x_j(1-x_j),$$

并将常数 K 用 C 表示出来.

以 $\sin\pi x_j$ 取代 $x_j(1-x_j)$ 后重复此分析过程.

第 7 章　线性代数方程组的迭代求解

在第 6 章，我们已经讨论了两种逼近线性椭圆型方程的方法，基于微分方程或其积分形式的有限差分法及有限元方法．每一种方法都给出一个规模可能很大的线性代数方程组．在实际的工程计算中，一个二维问题有数千个未知数，一个三维问题有数十万个未知数的情况都很常见．求解这样的方程组本身即为十分重要的问题，而且在这方面已有很多细致的研究．正如我们前面看到的，通过离散得到的方程组有很多特殊之处，一个高效的求解方法必须充分地利用这些特点．这些方程组最明显的性质就是它是极其稀疏的 (sparse), 即使有数千个未知数，每个方程也仅包含一个未知数以及与其紧邻的未知数．特别地，如果我们将方程组写为惯用的记法

$$A\mathbf{x} = \mathbf{b} \tag{7.1}$$

其中 A 是一个 $N \times N$ 的矩阵， $\mathbf{b}$ 是一个给定的数据向量， $\mathbf{x}$ 是关于 N 个内部未知网格点值的向量，那么这就意味着对这些值做了一个一维排序， 这多少有些不自然，而且掩盖了这样一个重要性质，即只有紧邻的点会包括在方程中．矩阵 A 的每一行只包括了很少的几个非零元素，一般是 5 个或 7 个；此外，对许多问题来说，适当排列未知数的次序，我们会得到一个非零元素规则排列的矩阵．这类结构上的性质，以及我们在第 6 章中强调的元素正负号的性质，在构造求解方程组的高效方法中都是重要的．

求解 (7.1) 的最明显的方法是直接的高斯消去法．对于自然排序的一维问题来说， 这是很有效的，用二阶精度的方法近似二阶方程，得到一个三对角方程组，而 2.9 节给出的 Thomas 算法与不选主元的高斯消去法是等价的；即便应用高阶格式，致使每个差分方程包含多于三个邻点，因而方程组将不再是三对角的，而是“带状”的 (即仅当 $|l-m|$ 较小时 $A_{lm} \neq 0$), 按照自然顺序的直接消去法效率也很高．

不过，在二维或三维中，任何顺序都会导致大得多的带宽． 譬如，考虑在覆盖了一致矩形网格的矩形区域上求解泊松方程，这样未知的内点值是 $U_{r,s}$, $r = 1, 2, \ldots, J_x$, $s = 1, 2, \ldots, J_y$. 我们按照标准排序或称 *自然排序*(natural ordering) 给诸点编号：从底部开始，每一行按照从左到右的顺序，依次逐行向上．那么在这个序列中未知数 $U_{r,s}$ 会编号为 $k = r + J_x(s-1)$. 那么，方程组中一个一般的方程会包含 $U_{r,s}$ 及其四个紧邻的网格点，它们是编号分别为 k, $k+1$, $k-1$, $k+J_x$, 和 $k-J_x$. 那么矩阵的非零元素都落在以对角线为中心的宽度为 $2J_x+1$ 的带状区域中．在解方程组的不选主元的消去过程中，带外的元素始终为零，但是带内则会充满非零元素．未知数的总数为 J_xJ_y, 而整个求解过程需要做的算术运算次数为 $J_xJ_y(2J_x+1)^2$ 量级，与处理非零矩阵所需的 $(J_xJ_y)^3$ 量级的运算相

比，这当然是少得多了，但仍很可观．对尺度为 h 的正方形网格，处理二维问题所需的算术运算次数为 h^{-4} 量级，对三维问题则为 h^{-7} 量级．对二维问题，消去法在处理任意区域与任意网格时的灵活性，已使其成为求解有限元方程优先采用的方法，但其统治地位已经受到了迭代法的挑战，并且广泛认同的看法是，对三维问题，迭代法的重要性会愈发增长．

在对稀疏矩阵使用直接消去法时，矩阵中被填入多少非零元素取决于矩阵各行的排序，对排序所产生的影响的研究已很成熟．对我们的模型问题来说，自然顺序绝非最佳，应用最优排序策略使得对规模极大的问题应用消去法成为可能，在 Duff, Erisman 和 Reid (1986) 的书中对用直接方法求解稀疏方程组有详细的描述．

迭代法最大限度地利用了矩阵 A 的结构．作为当时新兴的核能工业的部分成果，迭代法在 20 世纪 50 年代经历了快速的发展过程，为中子运输建模发展了有效的方法．在本章的第一部分，我们将对这些方法及其理论给出多少有些简化的介绍；接下来，在 7.6 节和 7.7 节，我们将介绍高效得多的现代方法，它们用到了这些基本的迭代过程．

7.1 显式基本迭代格式

假设我们将熟知的二维网格上的五点格式写为下列形式：

$$\tilde{c}_P U_P - [\tilde{c}_E U_E + \tilde{c}_W U_W + \tilde{c}_N U_N + \tilde{c}_S U_S] = b_P, \tag{7.2}$$

其中 $U_P, U_E, \ldots, U_S$ 是 U 的未知的 (内点) 值，b_P 代表所有已知数据，包括狄利克雷边界条件给出的边界值．这样，譬如 U_N 是边界上的已知值，那么这一项将会移到 (7.2) 的右端，它不会出现在左端，而 (7.2) 中的 $\tilde{c}_N$ 将为零．正是因为这一与第 6 章约定成俗的差别，我们在此采用了系数 $\tilde{c}_P$ 等而非 c_P 等；这个特点对应于有限元方法的实际实现过程，在计算如 (6.116) 中矩阵元素 A_{ij} 时，首先并不理会将要用到的边界条件，之后再将与已知的值 U_j 相关的部分移到右端．这种惯用的做法与第 5 章发展问题的做法也更加接近．

迭代的起点是对所有未知数的初始估计，我们用 $U_P^{(0)}, U_E^{(0)}, \ldots, U_S^{(0)}$ 来表示．最简单的迭代程序是雅可比在 1844 年首次使用的，通常称为**雅可比迭代法**(Jacobi method) 或**同时替换法** (method of simultaneous displacement)，它给出的逐次迭代值 $U_P^{(n)}$ 由下式定义：

$$U_P^{(n+1)} = (1/\tilde{c}_P)[b_P + \tilde{c}_E U_E^{(n)} + \tilde{c}_W U_W^{(n)} + \tilde{c}_N U_N^{(n)} + \tilde{c}_S U_S^{(n)}], \qquad n = 0, 1, 2, \ldots. \tag{7.3}$$

显然，此算法与未知数的排序无关，并且非常适用于现代的并行计算机．而且，如果所有的系数都是正的，而 $\tilde{c}_P$ 不小于其他系数之和，那么，如 6.3 节，从 (7.3) 中减去 (7.2) 的 $(1/\tilde{c}_P)$ 倍，便得到

$$|U_P^{(n+1)} - U_P| \leqslant \max_{E,W,N,S}\{|U_Q^{(n)} - U_Q|\}. \tag{7.4}$$

这还不足以证明迭代法的收敛性，因为证明收敛性需要不等式至少是严格的．不过，它起码说明这个序列不会无限发散，并且它体现了系数的上述性质的重要性．

然而，在一个串行计算机上，未知数 $U_P^{(n+1)}$ 的新值是按照特定次序由 (7.3) 计算的，而且使用最新得到的 $U_E, U_W, \ldots, U_S$ 的值似乎是有好处的：在任意时刻，计算机的内存中只需要对每个网格点保存一个值，而且似乎可以期待这将提高收敛速度．如果我们采用前面提到的自然顺序，那么 (7.3) 将替换为

$$U_P^{(n+1)} = (1/\tilde{c}_P)[b_P + \tilde{c}_E U_E^{(n)} + \tilde{c}_W U_W^{(n+1)} + \tilde{c}_N U_N^{(n)} + \tilde{c}_S U_S^{(n+1)}], \qquad n = 0, 1, 2, \ldots. \tag{7.5}$$

显然，在相同的条件下，我们将再次得到 (7.4)．这种迭代法通称为 *逐次替换法* (method of successive displacements) 或 *高斯 - 赛德尔方法* (Gauss-Seider method)．Gerling 在 1843 年宣称 Gauss 此前曾使用过该方法，而在 1874 年， Seider 又独立的发表了这个方法．有时也以曾经于 1918 年使用此方法的 Liebmann 来命名它．

当然了，这两种程序早期是人工实现的，并没有电子计算机的帮助．在逐次迭代中，(7.3) 或者 (7.5) 的规则用法常有变化以提高收敛速率：通常会监控方程的残量，可能会选择残量最大的来确定下一步需要更新的 U_P．20 世纪 40 年代，在 Southwell 的领导下，这种所谓的松弛方法在牛津这里达到了极高的发展水平．一个成果就是高斯 - 赛德尔程序的修正，现在称为 *逐次超松弛* (successive over-relaxation) 或 SOR *方法*(SOR method)．将 U_P 的新值取为旧值和由 (7.5) 给出的值的加权平均，权为 ω 和 $1-\omega$, 我们便得到

$$\begin{aligned} U_P^{(n+1)} &= (1-\omega)U_P^{(n)} + (\omega/\tilde{c}_P)[b_P + \tilde{c}_E U_E^{(n)} + \tilde{c}_W U_W^{(n+1)} + \tilde{c}_N U_N^{(n)} + \tilde{c}_S U_S^{(n+1)}] \\ &= U_P^{(n)} + (\omega/\tilde{c}_P)[b_P + \tilde{c}_E U_E^{(n)} + \tilde{c}_W U_W^{(n+1)} + \tilde{c}_N U_N^{(n)} + \tilde{c}_S U_S^{(n+1)} - \tilde{c}_P U_P^{(n)}] \\ &= U_P^{(n)} - (\omega/\tilde{c}_P) r_P^{(n)}, \end{aligned} \tag{7.6}$$

其中 $r_P^{(n)}$ 是点 P 处的残量；这个残量是方程左右两端的差，并由未知数最新得到的值计算得出．表示同一种程序的另一种做法是，计算高斯 - 赛德尔迭代 (7.5) 给出的修正，乘以 ω 后再加到以前的值上． *超松弛* (over-relaxation) 的说法隐含了 $\omega > 1$．

在最后一个公式 (7.6) 中，我们可以立刻看出它与求解相应的抛物型方程格式之间的类似之处：如果用雅可比方法，则其与通常的显式格式是完全相同的．我们可以令 ω 等于 $4\Delta t/(\Delta x)^2$, 并可以期待 ω 取到稳定性所允许的最大值时得到最好的收敛性．在雅可比方法下就是 $\omega = 1$: 但是我们可以预测在 (7.6) 中，它将会因更多的隐式性质而增大，因此超松弛就对应 $\omega > 1$; $\omega < 1$ 时称为低松弛 (under-relaxation), 而 $\omega = 1$ 对应高斯 - 赛德尔迭代．注意，对此方法，超松弛因子 ω 是取为常数的，不仅对于每个未知数，而且对于每

一步迭代都使用相同的因子值.

7.2 迭代法的矩阵形式及其收敛性

按照 (7.1) 的记法，我们把方程组的矩阵写成如下形式:

$$A = D - L - U \tag{7.7}$$

其中 D 是对角阵，对角元素都是严格正数 (对应于系数 $\tilde{c}_P$), L 是严格下三角阵，而 U 是严格上三角阵. 矩阵 L 和 U 的前面写上负号，是因为在第 6 章的问题中，对角线以外的矩阵元素都是非正的. 未知数的排序完全决定了差分方程的哪些系数出现在 L 中，哪些出现在 U 中. 这样我们可以把此前介绍的三种格式写为:

雅可比:

$$\mathbf{x}^{(n+1)} = D^{-1}[\mathbf{b} + (L+U)\mathbf{x}^{(n)}]; \tag{7.8}$$

高斯 - 赛德尔:

$$D\mathbf{x}^{(n+1)} = \mathbf{b} + L\mathbf{x}^{(n+1)} + U\mathbf{x}^{(n)}, \tag{7.9a}$$

即

$$\mathbf{x}^{(n+1)} = (D-L)^{-1}[\mathbf{b} + U\mathbf{x}^{(n)}], \tag{7.9b}$$

或

$$\mathbf{x}^{(n+1)} = (I - D^{-1}L)^{-1}[D^{-1}\mathbf{b} + D^{-1}U\mathbf{x}^{(n)}]; \tag{7.9c}$$

SOR:

$$D\mathbf{x}^{(n+1)} = (1-\omega)D\mathbf{x}^{(n)} + \omega[\mathbf{b} + L\mathbf{x}^{(n+1)} + U\mathbf{x}^{(n)}], \tag{7.10a}$$

即

$$\mathbf{x}^{(n+1)} = (D-\omega L)^{-1}[\omega\mathbf{b} + \omega U\mathbf{x}^{(n)} + (1-\omega)D\mathbf{x}^{(n)}], \tag{7.10b}$$

或

$$\mathbf{x}^{(n+1)} = (I - \omega D^{-1}L)^{-1}[\omega D^{-1}\mathbf{b} + \{\omega D^{-1}U + (1-\omega)I\}\mathbf{x}^{(n)}]. \tag{7.10c}$$

这些形式对下面的分析是有用的，但是前面的显示格式更清楚地体现了每个算法究竟是如何实现的.

以上的每种迭代方法都可以写为

$$\mathbf{x}^{(n+1)} = G\mathbf{x}^{(n)} + \mathbf{c}, \tag{7.11}$$

其中 G 称为该方法的迭代矩阵 (iteration matrix). 迭代矩阵 G 可由 (7.8), (7.9) 或者 (7.10) 毫无困难地得到，方程组的解 $\mathbf{x}$ 满足

$$(I-G)\mathbf{x}=\mathbf{c}. \tag{7.12}$$

这样，如果 $\mathbf{e}^{(n)}:=\mathbf{x}^{(n)}-\mathbf{x}$ 是 n 次迭代之后的误差，则

$$\mathbf{e}^{(n+1)}=G\mathbf{e}^{(n)}=G^2\mathbf{e}^{(n-1)}=\cdots=G^{n+1}\mathbf{e}^{(0)}. \tag{7.13}$$

引理 7.1. 迭代 (7.11)-(7.13) 对任意的起始向量 $\mathbf{x}$ 随 $n\to\infty$ 都收敛，当且仅当

$$\rho(G)<1, \tag{7.14}$$

其中 $\rho(G)$ 是 G 的谱半径 (spectral radius), 即 $\rho(G)=\max\limits_i|\lambda_i(G)|$. 其渐近收敛速率 (asymptotic rate of convergence) 为 $-\ln\rho$.

证明. 对于所有初始向量都成立的收敛性，其充要条件是随 $n\to\infty$, $\|G^n\|\to 0$. 首先假设矩阵 G 的所有特征向量张成其所在的欧氏空间，那么存在非奇异矩阵 S, 使得

$$G=S\Lambda S^{-1} \tag{7.15}$$

其中 Λ 是一个对角矩阵，对角元素是矩阵 G 的特征值 $\{\lambda_i\}$. 则

$$G^n=S\Lambda^nS^{-1}. \tag{7.16}$$

矩阵 Λ^n 也是一个对角阵，其对角元素为 λ_i^n, 它们趋于零，当且仅当每个 $|\lambda_i|<1$.

如果 G 没有上述性质，它也一定可以写为类似 (7.15) 的形式，只是 Λ 由 G 的若当形 (Jordan form) J 取代. 论证本质是类似的. ■

收敛速率定义为 $R=-\ln\rho$, 显示了为达到要求的精度需要的迭代次数. 譬如，为使误差缩减 10^p 倍，需要的迭代步数约为 $(p/R)\ln 10$. 严格说来，R 称为渐近收敛速率，因为它仅对充分大的 n 精确显示所需的迭代步数；在迭代的初期，误差的缩减会比其显示的快些或慢些. 这可由 (7.13) 和 (7.15) 得出，因为

$$\|\mathbf{e}^{(n)}\|\leqslant\|G^n\|\|\mathbf{e}^{(0)}\|\leqslant\|S\|\|S^{-1}\|\|\Lambda\|^n\|\mathbf{e}^{(0)}\|, \tag{7.17a}$$

即

$$(\|\mathbf{e}^{(n)}\|/\|\mathbf{e}^{(0)}\|)^{1/n}\leqslant(\|S\|\|S^{-1}\|)^{1/n}\rho; \tag{7.17b}$$

而对属于最大特征值的特征向量来说，误差的缩减就恰为 ρ.

将矩阵对角占优的概念和第 6 章中用于分析差分格式的最大值原理形式化地联系起来是有用的. 一个矩阵称为是对角占优的 (diagonally dominant), 如果其每一行，在绝对值意义下，对角元素都不小于非对角元素的和；即

$$|A_{ll}|\geqslant\sum_{m\neq l}|A_{lm}|,\qquad\forall l. \tag{7.18}$$

这是差分格式的条件 (6.44) 所隐含的. 如果对 (7.18) 中的所有 l, 严格不等式都成立, 则称该矩阵是*严格对角占优的* (strictly diagonally dominant). 于是很容易证明雅可比迭代法对严格对角占优的矩阵收敛. 不过, 这个结论没什么太大的帮助, 因为我们的矩阵几乎都不满足这一条件.

一个矩阵称为是*不可约对角占优的* (irreducibly diagonally dominant), 如果 (7.18) 成立, 且其中对至少一个(at least one) l 严格不等式成立, 同时它是*不可约的*(irreducible). 不可约性的概念与第 6 章中节点集合的连通性质是紧密关联的. 如果元素 A_{lm} 非零, 则称矩阵的第 l 行和第 m 行是*连通的*(connected). 那么, 一个矩阵是*不可约的*(irreducible), 如果对任意的两行 l_l 和 l_k, 总存在一个序列 $l_1, l_2, \ldots, l_k$, 使得序列中的每一行总与其下一行连通. 很明显, 建立在连通点集上的差分格式生成的是一个不可约矩阵; 此外, 必须在至少一行成立严格不等式的条件与早些时候提出的必须在至少一部分边界上给出狄利克雷边界条件的条件相对应. 因此, 这些条件都包含在第 6 章中提出的使最大值原理成立的那些条件中.

定理 7.1. *雅可比迭代法对不可约对角占优矩阵是收敛的.*

证明. 假设我们以 μ 表示雅可比迭代矩阵的任意特征值, $\mathbf{v}$ 是相应的特征向量; 那么我们必须证明 $|\mu| < 1$, 这里严格不等式是必须的. 我们有

$$D^{-1}(L+U)\mathbf{v} = \mu\mathbf{v}, \tag{7.19a}$$

或

$$(\mu D - L - U)\mathbf{v} = \mathbf{0}. \tag{7.19b}$$

我们构建最大值原理, 方法与用于引理 6.1 的证明的类似. 假设 $|\mu| = 1$ 且 v_k 是特征向量 $\mathbf{v}$ 的按模最大的分量. 那么我们得到

$$\mu A_{kk} v_k = -\sum_{m \neq k} A_{km} v_m, \tag{7.20}$$

由此推出

$$|\mu| \leqslant \sum_{m \neq k} \frac{|A_{km}|}{|A_{kk}|} \frac{|v_m|}{|v_k|}. \tag{7.21}$$

这样如果 $|\mu| = 1$, 我们就一定会得到在对角占优关系 (7.18) 中, $l = k$ 时等式成立, 同时对所有与 k 连通的行 m, 有 $|v_m| = |v_k|$. 用每一个 m 来代替 k, 重复以上论证, 递推地重复这个过程, 由不可约假设, 我们可以取遍所有的行. 但是严格不等式须对 (7.18) 中的至少一个 l 值成立. 所以 $|\mu| = 1$ 的假设导致了矛盾. ∎

我们还可以给出 SOR 方法中的松弛因子 ω 的界.

定理 7.2. 如果 $0 < \omega < 2$ 不成立，则 SOR 方法不收敛.

证明. SOR 迭代矩阵的特征值满足

$$\det[(D-\omega L)^{-1}(\omega U+(1-\omega)D)-\lambda I]=0, \tag{7.22a}$$

或等价地

$$\det[\lambda(D-\omega L)-\omega U-(1-\omega)D]=0, \tag{7.22b}$$

或

$$\det[(\lambda+\omega-1)I-\lambda\omega D^{-1}L-\omega D^{-1}U]=0. \tag{7.22c}$$

若展开这个行列式，我们将得到一个关于 λ 的多项式，其主项和常数项仅来自对角元素. 多项式具有如下形式，

$$\lambda^N+\cdots+(\omega-1)^N=0, \tag{7.23}$$

其中 N 是矩阵的阶. 因此特征值的乘积为 $(1-\omega)^N$，这样如果 $|1-\omega|\geqslant 1$，那么至少一个特征值会满足 $|\lambda|\geqslant 1$，因而迭代不收敛. 注意这个结果对矩阵没有做任何假设，它可以是很一般的结论. ■

这里值得强调的是，本章我们所用的收敛概念与第 2 至 5 章，还有第 6 章所用到的意义截然不同. 在所有那些章中，我们关注的是当网格加密时，U 趋向于 u 的收敛性：在收敛过程的每一阶段，我们有一个有限维近似，但是，随着网格加密，维数无限制增长. 而这里我们有一个固定网格，和一个严格有限维的用迭代法求解的代数问题：只有当这个问题收敛时，我们才真正地得到了 u 在这个网格上的逼近 U. 加密网格以使 U 收敛于 u，引进了也可以用迭代法求解的更复杂的代数问题. 这样我们所要讨论的收敛过程本身是置于一个更大的收敛过程的框架下的. 明辨这些区别是很重要的，因为，正如已经说过的，有些迭代方法与用于求解抛物型方程的发展格式是类似的. 这反映在 $U^{(n)}$ 等符号上：这里我们令 $n\to\infty$ 时，Δx 是固定的；而在前几章中，我们有 $\Delta t,\Delta x\to 0$，而 $n\to\infty$ 只是由此产生的结果.

7.3 收敛性的傅里叶分析

对特别的模型问题，我们可以利用傅里叶分析来更精确地查验各种迭代法的收敛速率. 假设我们要求解逼近单位正方形上赋狄利克雷边界条件的 $\nabla^2 u+f=0$ 的五点差分格式，对一个尺寸为 Δx(满足 $J\Delta x=1$) 的正方形网格，我们有 $N=(J-1)^2$ 个未知的内点值. 误差可以按傅里叶波型展开，对于这里考虑的赋狄利克雷边界条件的情形，仅用到正弦波型；在网格点 (x_r,y_s) 处，我们有

$$e_{r,s}^{(n)} = \sum_{k_x,\,k_y} a^{(n)}(k_x,\ k_y) \sin k_x x_r \sin k_y y_s, \tag{7.24a}$$

其中

$$k_x,\ k_y = \pi,\ 2\pi, \ldots, (J-1)\pi. \tag{7.24b}$$

在 (7.3) 中，取 $\tilde{c}_P = 4, \tilde{c}_E = \tilde{c}_W = \tilde{c}_N = \tilde{c}_S = 1$, 应用雅可比方法，由 $\sin A + \sin B$ 的三角公式，显然有

$$e_{r,s}^{(n+1)} = \sum_{k_x,\,k_y} a^{(n)}(k_x,\ k_y)\frac{1}{4}(2\cos k_x\Delta x + 2\cos k_y\Delta x) \sin k_x x_r \sin k_y y_s. \tag{7.25}$$

因此 $\sin k_x x_r \sin k_y y_s$ 就是雅可比迭代矩阵 $G_J := D^{-1}(L+U)$ 的一个特征向量，其从属的特征值是

$$\mu_J(k_x, k_y) = \frac{1}{2}(\cos k_x\Delta x + \cos k_y\Delta x). \tag{7.26}$$

取遍 (7.24b) 所列出值就给出了全部特征向量；因此 $|\mu_J|$ 的最大值出现在 $k_x = k_y = \pi$ 或 $(J-1)\pi$ 的极端情况．在这种情况下 $\mu_J = \pm\cos\pi\Delta x$, 于是

$$\max|\mu_J| = \cos\pi\Delta x \sim 1 - \frac{1}{2}(\pi\Delta x)^2 + \cdots. \tag{7.27}$$

这显然意味着当 Δx 值很小时, 收敛会非常慢, 并且那些频率最低的波型, 即每个方向上的 $\sin j\pi\Delta x$ 收敛得最慢；这种情况也出现在具有形式 $\sin j(J-1)\pi\Delta x = (-1)^{j-1}\sin j\pi\Delta x$ 的频率最高的波型．不过，在后一种情况，我们有 $\mu_J \sim -1$, 因此这样的误差波型很容易通过对相继的两步迭代取平均来减弱．

收敛速率是

$$\begin{aligned} -\ln|\mu_J| &= -\ln(\cos\pi\Delta x) \\ &\sim -\ln(1 - \frac{1}{2}(\pi\Delta x)^2) \\ &\sim \frac{1}{2}(\pi\Delta x)^2. \end{aligned} \tag{7.28}$$

例如，取网格尺寸为 $\Delta x = 0.02$, 上式值为 0.00197, 这样为使误差缩减 10 倍需要 1166 步迭代．如果估计初始误差的阶为 1, 而我们希望找到有六位 10 进制小数精度的解，那么需要约 7000 步迭代．因此对其他可能收敛的更快些的方法的研究就显得重要了．

可是，对于 SOR 迭代法 (7.6), 情况就复杂得多了，这是因为 $\sin k_x x_r \sin k_y y_s$ 不再是迭代矩阵的特征向量．正如稳定性分析的做法，处理复形式要更方便些．所以我们假设要求具有形式

$$e_{r,s}^n = [g(\xi,\eta)]^n e^{\mathrm{i}\,(\xi r + \eta s)} \tag{7.29}$$

的误差，其中 $\xi = k_x\Delta x$, $\eta = k_y\Delta y$. 将其代入 (7.6) 可以看出，当

$$g(\xi,\eta) = \frac{1-\omega+\frac{1}{4}\omega(\mathrm{e}^{\mathrm{i}\,\xi}+\mathrm{e}^{\mathrm{i}\,\eta})}{1-\frac{1}{4}\omega(\mathrm{e}^{-\mathrm{i}\xi}+\mathrm{e}^{-\mathrm{i}\,\eta})} \tag{7.30}$$

时，这种形式的误差为一个特征向量. 不难看出 $g(-\xi,-\eta) \neq g(\xi,\eta)$[1], 这也再次说明为什么 $\sin\xi r\sin\eta s$ 不是迭代矩阵的特征向量.

不过，利用 Garabedian 首先提出，并为 Leveque 和 Trefethen[2] 后继发展起来的斜置时间 -空间网格的方法会改善此局面. 我们引入新的指标和特征向量的形式

$$\nu = 2n + r + s, \tag{7.31a}$$

$$e^{\nu}_{r,s} = [g(\xi,\eta)]^{\nu}\mathrm{e}^{\mathrm{i}\,(\xi r+\eta s)}. \tag{7.31b}$$

在新的指标下，相应于 (7.6) 的误差方程涉及到三层，具有如下形式

$$e^{\nu+2}_{r,s} = (1-\omega)e^{\nu}_{r,s} + \frac{1}{4}\omega[e^{\nu+1}_{r+1,s} + e^{\nu+1}_{r-1,s} + e^{\nu+1}_{r,s-1} + e^{\nu+1}_{r,s+1}]. \tag{7.32}$$

将 (7.31b) 代入得到一个关于 $g(\xi,\eta)$ 的二次方程

$$g^2 = (1-\omega) + \frac{1}{2}w[\cos\xi + \cos\eta]g, \tag{7.33}$$

其中各系数是 ξ 和 η 的偶函数. 因此迭代矩阵的特征向量就有如下的形式

$$[g(\xi,\eta)]^{r+s}\sin\xi r\sin\eta s, \tag{7.34}$$

且其最大的特征值是

$$\max|\lambda_{\mathrm{SOR}}| = \max_{\xi,\eta}\max\{|g_+(\xi,\eta)|^2, |g_-(\xi,\eta)|^2\}, \tag{7.35}$$

其中 g_+ 和 g_- 是 (7.33) 的根.

现在就清楚了，这个结果实际上可以通过直接查验具有形式

$$\lambda^n[g(\xi,\eta)]^{r+s}\sin\xi r\sin\eta s \tag{7.36}$$

的误差的性态得到. 将其代入 (7.6), 即知我们要求

$$\begin{aligned}\lambda\sin\xi r\sin\eta s \quad = \quad & 1-\omega+\frac{1}{4}w[g\sin\xi(r+1)\sin\eta s + (\lambda/g)\sin\xi(r-1)\sin\eta s \\ & +g\sin\xi r\sin\eta(s+1) + (\lambda/g)\sin\xi r\sin\eta(s-1)].\end{aligned} \tag{7.37}$$

因此只要 $g = \lambda/g$, 我们即得到正确的三角和，而且在这种情况下，我们便得到了一个特征向量，其相应的特征值是

$$\lambda = 1-\omega+\frac{1}{2}\omega(\cos\xi+\cos\eta)\lambda^{1/2}. \tag{7.38}$$

---

[1] 原文为 $g(\pm\xi,\pm\eta) \neq g(\xi,\eta)$

[2] LeVeque, R.J. and Trefethen, L.N.(1988), Fourier analysis of the SOR iteration, *IMA J. Numer. Anal.* **8**(3), 273-9.

这与 (7.33) 吻合.

$\omega = 1$ 的高斯 - 赛德尔迭代是一种特殊情况; 由于那时 (7.33) 的一个根是 $g = 0$, 于是 0 是一个多重特征值, 从而矩阵的特征向量不能张成其所在的欧氏空间. 不过我们可以在 (7.38) 式中令 $\omega \to 1$ 取极限, 并由此推出高斯 - 赛德尔迭代矩阵的特征值是 0 和雅可比迭代矩阵的特征值的平方. 这说明了高斯 - 赛德尔方法收敛速度恰是雅可比方法的两倍.

更一般地, 当 $\omega \neq 1$, 由 (7.33) 容易证实, $0 < \omega < 2$ 对 $|g_\pm| < 1$ 是必要的. 进一步说, 鉴于 $|g_\pm|$ 关于 $\mu = \frac{1}{2}(\cos\xi + \cos\eta)$ 的依赖性, 容易验证其最大值取在 $\mu_0 = \cos\pi\Delta x$; 再考察最大值关于 ω 的变化规律, 我们可以找到最优的 SOR 因子.

SOR 矩阵的特征值满足 (7.38). 对固定的 ω 的值, 最大的特征值对应于 $\frac{1}{2}(\cos\xi + \cos\eta) = \mu_0 = \cos\pi\Delta x$. 于是方程 (7.38) 可以写成

$$\lambda^2 + 2\lambda(\omega - 1 - \frac{1}{2}\mu_0^2\omega^2) + (\omega - 1)^2 = 0. \tag{7.39}$$

显然, 当 $\mu_0^2\omega^2 \leqslant 4(\omega - 1)$ 时, 这只可能对 $\omega > 1$ 成立, 它们形成了模等于 $\omega - 1$ 的复共轭对; 而在相反的情况下, 由 (7.39) 我们看出这时根的和是正数, 因此它们两个都是正的. 我们现在容易得到

$$\lambda_{\text{SOR}} = 1 - \omega + \frac{1}{2}\mu_0^2\omega^2 + \mu_0\omega(1 - \omega + \frac{1}{4}\mu_0^2\omega^2)^{1/2}, \quad 若\omega < 1 + \frac{1}{4}\mu_0^2\omega^2, \tag{7.40a}$$

$$\lambda_{\text{SOR}} = \omega - 1, \qquad 若\omega \geqslant 1 + \frac{1}{4}\mu_0^2\omega^2. \tag{7.40b}$$

对一个典型的 μ_0 值, 其性态如图 7-1 所示. 由于 (7.40a) 的表达式关于 ω 递减, 而

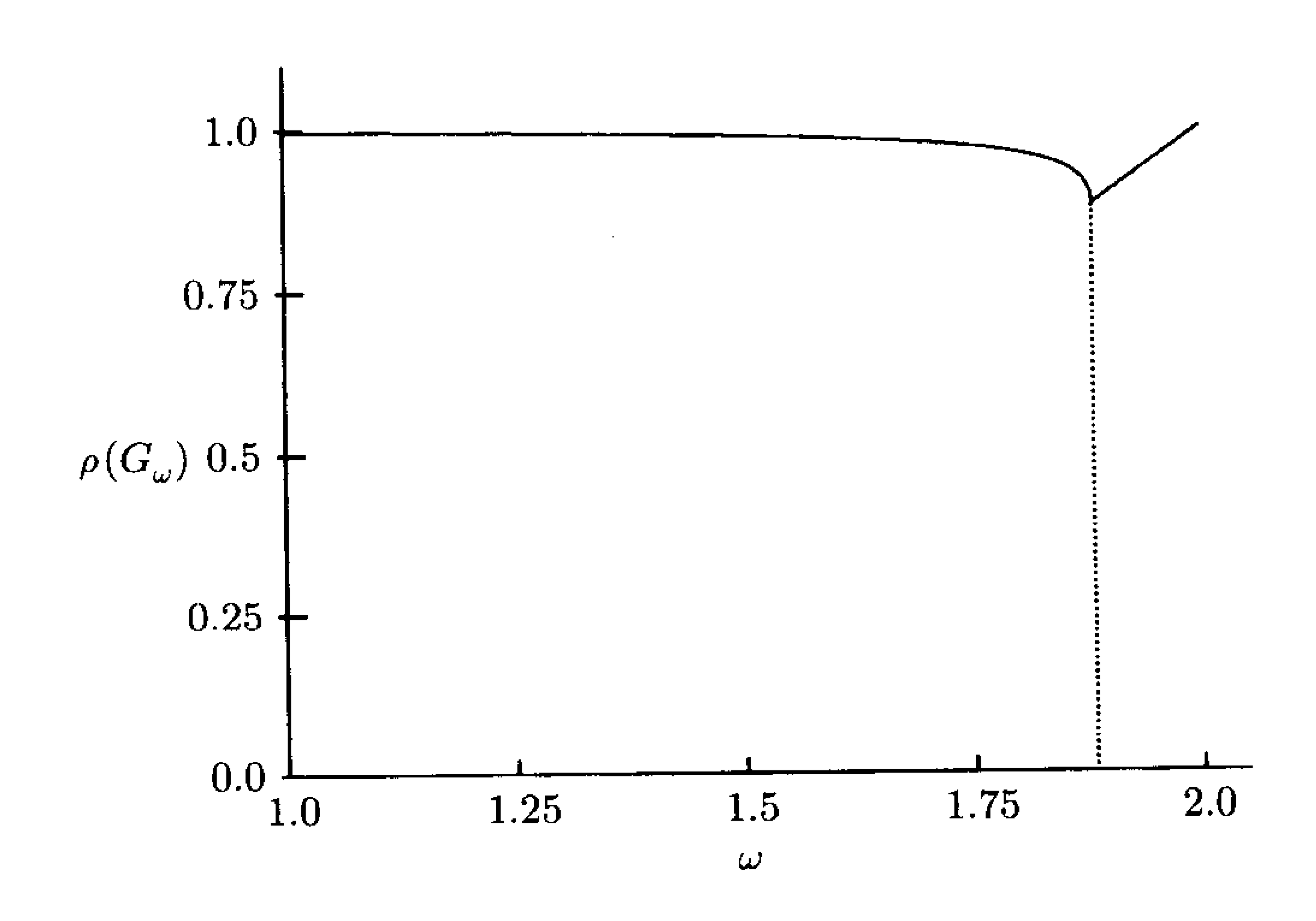


图 7-1 对 $\mu_0 = 0.99$, 谱半径 $\rho(G_\omega)$ 关于 ω 的函数

(7.40b) 则递增，所以最优值出现在 $\omega = 1 + \frac{1}{4}\mu_0^2\omega^2$, 即

$$\omega = \frac{2}{\mu_0^2}(1 - \sqrt{1 - \mu_0^2}). \tag{7.41}$$

代入 μ_0 的值，我们得到

$$\begin{aligned} \omega_{opt} &= \frac{2}{1 + \sin \pi \Delta x} \sim 2 - 2\pi\Delta x, \\ \rho(G_\omega)_{opt} &= \frac{1 - \sin \pi \Delta x}{1 + \sin \pi \Delta x} \sim 1 - 2\pi\Delta x. \end{aligned} \tag{7.42}$$

这样，由 $-\ln \rho(G_\omega)$ 给出的渐近收敛速度近似为 $2\pi\Delta x$. 当 $\Delta x = 0.02$ 时，它是 0.1256, 所以要使误差缩小 10 倍仅需 18 步迭代；与此相比雅可比迭代法需要 1166 步（见 7.3 节）.

7.4 应用于一个例子

应用最佳松弛因子给收敛速度带来了非常明显的改善. 在模型问题中，我们可以计算雅可比矩阵的特征值，从而找出最佳因子. 而在更一般的问题中，这种做法是不可行的，因此有必要选择某个有意义的 ω. 图 7-2 在 $1 < \omega < 2$ 的范围内画出了模型问题作

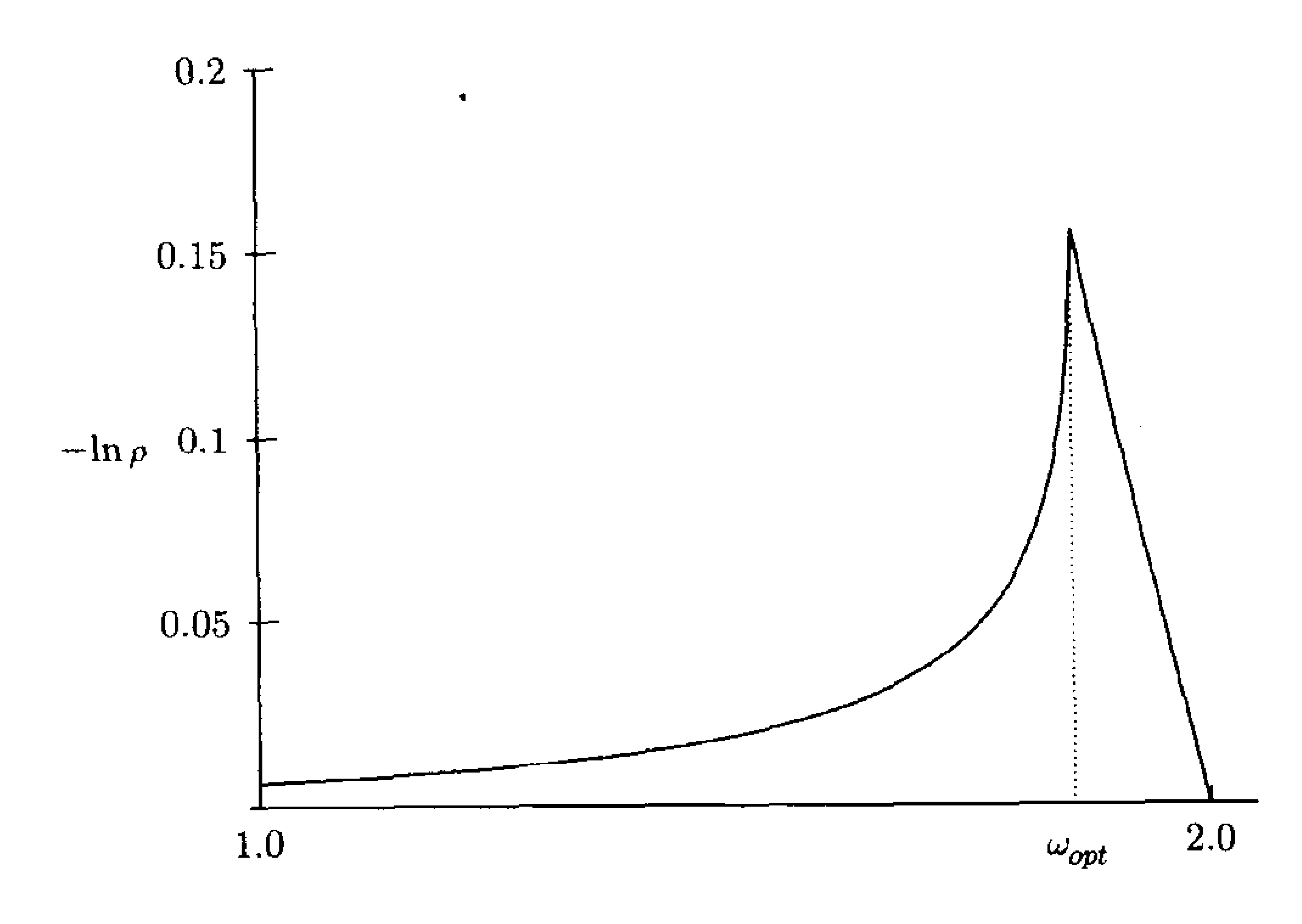


图 7-2 SOR 迭代的收敛速度关于 ω 的函数

为 ω 的函数的收敛速度. 通过这个相当典型的例子清楚地看出，最优值处的变化十分陡峭，任何偏差都会很不幸地导致收敛速度的迅速下降. 不过，它确实也反映出高估因子总要强于将因子低估. 当 $\omega > \omega_{opt}$, 特征值是复数，其模是 $\omega - 1$; 图 7-2 中曲线的斜率接近于 -1. 但是在临界值的左边，斜率趋于无穷，ω 很小的偏差都会带来很大的影响.

作为一个关于 SOR 迭代的性态的例子，我们应用此方法求解在第 6 章末尾所讨论的问题所产生的线性方程组. 此处，我们采用了标准的五点差分格式近似一致正方形网格

上的拉普拉斯方程. 严格地讲, 7.3 节的傅里叶分析在此并不适用, 因为 (7.24a) 的特征向量在曲线边界上不满足正确的边界条件. 我们或可期望边界的影响会相当小, 从而 (7.42) 给出的因子选择虽然不是真正最优的, 但是仍会给出好的收敛性.

计算中采用区间 $\Delta x = 0.025$, 圆形边界的半径为 0.4. 迭代起始于内点处的值均为 0 的初始向量. 图 7-3 画出了对不同的松弛因子, 迭代过程中的误差性态. 为确定每个阶段的误差, 我们首先进行了很多次数的迭代, 以使序列收敛到代数方程组的真解 $\{U_{r,s}\}$. 在每步迭代后, 误差通过 $E^{(n)}$ 衡量, 这里

$$E^{(n)} = \left[\frac{1}{N}\sum_{r,s}(U_{r,s}^{(n)} - U_{r,s})^2\right]^{1/2}. \tag{7.43}$$

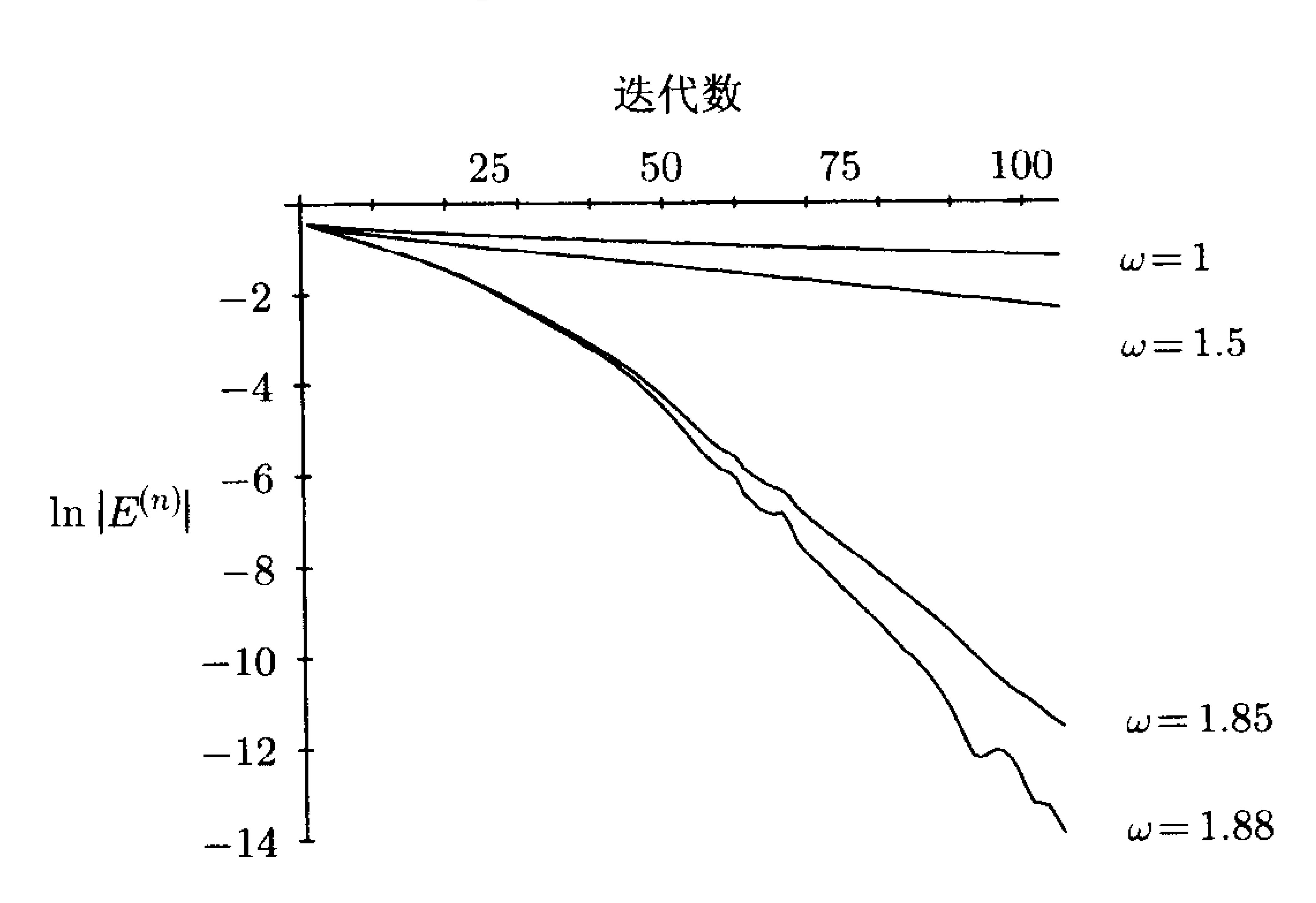


图 7-3 SOR 迭代的收敛性

图像将 $E^{(n)}$ 表示为 n 的函数, 并且很清楚地体现了收敛性较之 $\omega = 1$ 的高斯–赛德尔迭代的巨大改进. $\Delta x = 0.025$ 时, 通过 (7.42) 预测所选取的最佳因子是 1.8545, 这带来了好的收敛性. 不过, 我们发现, 一个稍微不同的值 $\omega = 1.88$ 就带来了显著的改进.

我们采用迭代矩阵的谱半径来衡量收敛速度, 并不总能对实际情况做出准确地描述. 如果我们将误差按矩阵的特征向量展开, 如 (7.24a) 所做, 每迭代一步, 各项的贡献都应该乘以相应的特征值. 如果最大的特征值显著地大于其他的特征值, 那么它的作用将最终会决定误差的性态. 但是在实际问题中, 更可能出现的是若干特征值具有几乎相同的模, 并且它们都会对误差有显著影响. 当选择 ω 接近最优值时, 这些特征值很可能都是复数, 这导致了图 7-3 中所示的不规则的性态. 对于小一些的 ω 值, 大部分特征值是实数, 因而误差性态表现得更为光滑.

7.5 推广及相关的迭代法

前几章的处理方法经常可以通过将矩阵 A 分块而改进收敛性，代价是增加一些额外的运算量. 假设 A 按如下形式分块

$$A = \begin{pmatrix} D_1 & -U_{12} & -U_{13} & \cdots \\ -L_{21} & D_2 & -U_{23} & \cdots \\ -L_{31} & -L_{32} & D_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \tag{7.44}$$

其中矩阵 D_k 是方阵，但大小不必相同. 则分块 SOR 方法的定义与 (7.10a) 类似，只是此时的 D, L 和 U 为矩阵

$$D = \begin{pmatrix} D_1 & 0 & 0 & \cdots \\ 0 & D_2 & 0 & \cdots \\ 0 & 0 & D_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \tag{7.45}$$

$$L = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ L_{21} & 0 & 0 & \cdots \\ L_{31} & L_{32} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}, \tag{7.46}$$

$$U = \begin{pmatrix} 0 & U_{12} & U_{13} & \cdots \\ 0 & 0 & U_{23} & \cdots \\ 0 & 0 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}. \tag{7.47}$$

与前面讨论的点迭代法的实际差别在于，现在由 (7.10a) 计算 $\mathbf{x}^{n+1}$ 需要求解一组相互分离的线性方程组，系数矩阵为 $D_1, D_2, D_3, \cdots$. 在我们的模型问题中，矩阵 A 的阶是 $(J-1)^2$；按自然顺序给未知数编号，我们可以将 A 分成若干块，每一个 D_k 都是 $(J-1)\times(J-1)$ 矩阵. 所有的矩阵 L_{lm} 和 U_{lm} 在 $|l-m|=1$ 时是对角阵，否则是零矩阵. 对角线上的分块矩阵 D_k 是三对角的，所以计算 $\mathbf{x}^{n+1}$ 需要求解 $J-1$ 个线性方程组，每个都是 $J-1$ 阶三对角的；这些方程组每个都可以用 Thomas 算法高效解出.

我们还可以按照列的次序给未知数编号，每列从底到顶排序. 将新的矩阵分成同样大小的块给出了新的块迭代法，这种方法是按照网格的竖直的块而非水平的块处理未知数. 如果我们采用一次水平线迭代紧跟一次竖直线迭代所合成的迭代法，则其一步迭代

的结果与 3.2 节中用于热传导方程 $u_t = \nabla^2 u$ 的 Peaceman-Rachford ADI 方法的一个时间步类似．如果我们在 (7.10a) 中取 $\omega = 1$, 在 (3.15) 中取 $\nu_x = \nu_y = 1$, 并在 (7.10a) 中用 $L\mathbf{x}^{(n)}$ 取代 $L\mathbf{x}^{(n+1)}$, 则两者就完全一致了．最后这种修正将基于高斯 - 赛德尔迭代的方法改为了基于雅可比迭代的方法．

这种交替线迭代方法与求解依赖时间的热传导方程，并寻求其长时间后的定常解的 ADI 方法紧密相关．但是迭代方法更为一般，因为我们并不需要迭代过程中各阶段与热传导方程各特定时间的解精确吻合，而只需要极限解相吻合．

不过，在许多实际应用领域，我们到现在为止所讨论的这些方法大半已被更有效且通常更精巧的方法所取代．在本章的最后，我们简要地介绍 *多重网格法* (multigrid method) 和 *共轭梯度法* (conjugate gradient method); 这两种方法和它们的各种变形是现在实际计算中最常用到的方法．

7.6　多重网格法

这个最早由 Brandt[1] 提出的方法，在过去 20 年已经发展成为一个求解由偏微分方程导出的代数方程组的非常高效的方法．在大多数情况下，这种高效源自其对原微分方程的利用．介绍其基本思想的最好方法是通过一个例子来说明．我们用到的是一个涉及自伴方程的例子：

$$(a(x,y)u_x)_x + (a(x,y)u_y)_y + f(x,y) = 0, \qquad (x,y) \in \Omega, \tag{7.48a}$$

$$u = 0, \qquad (x,y) \in \partial\Omega, \tag{7.48b}$$

其中 Ω 是单位正方形

$$\Omega := (0,1) \times (0,1), \tag{7.49}$$

而 $\partial\Omega$ 是此正方形的边界．函数 a 为

$$a(x,y) = x + 2y^2 + 1, \tag{7.50}$$

并且选择 f 以使差分方程的解给出

$$u(x,y) = xy(1-x)(1-y) \tag{7.51}$$

在网格点上的逼近．

采用中心差分格式 (6.5), 先取 $\Delta x = 1/10$, 然后取 $\Delta x = 1/20$. 我们首先用高斯 - 赛德尔迭代求解由此得到的两个方程组．为了凸显多重网格方法思想的基础，我们从一个由其各元素是在 $[0,1]$ 上均匀分布的随机数组成的初始向量开始迭代，尽管在实际应用中这

[1] Brandt, A.(1977) Multi-level adaptive solutions to boundary value problems. *Math. Comput.* **31**, 333-90.

并不是一个有用的初始向量. 迭代结果如图 7-4 所示. 图像显示的是 $\ln(R_n)$, 这里 R_n 是 n 步迭代后残量向量的 2- 范数; 实线相应于 $\Delta x = 1/10$, 点线对应 $\Delta x = 1/20$.

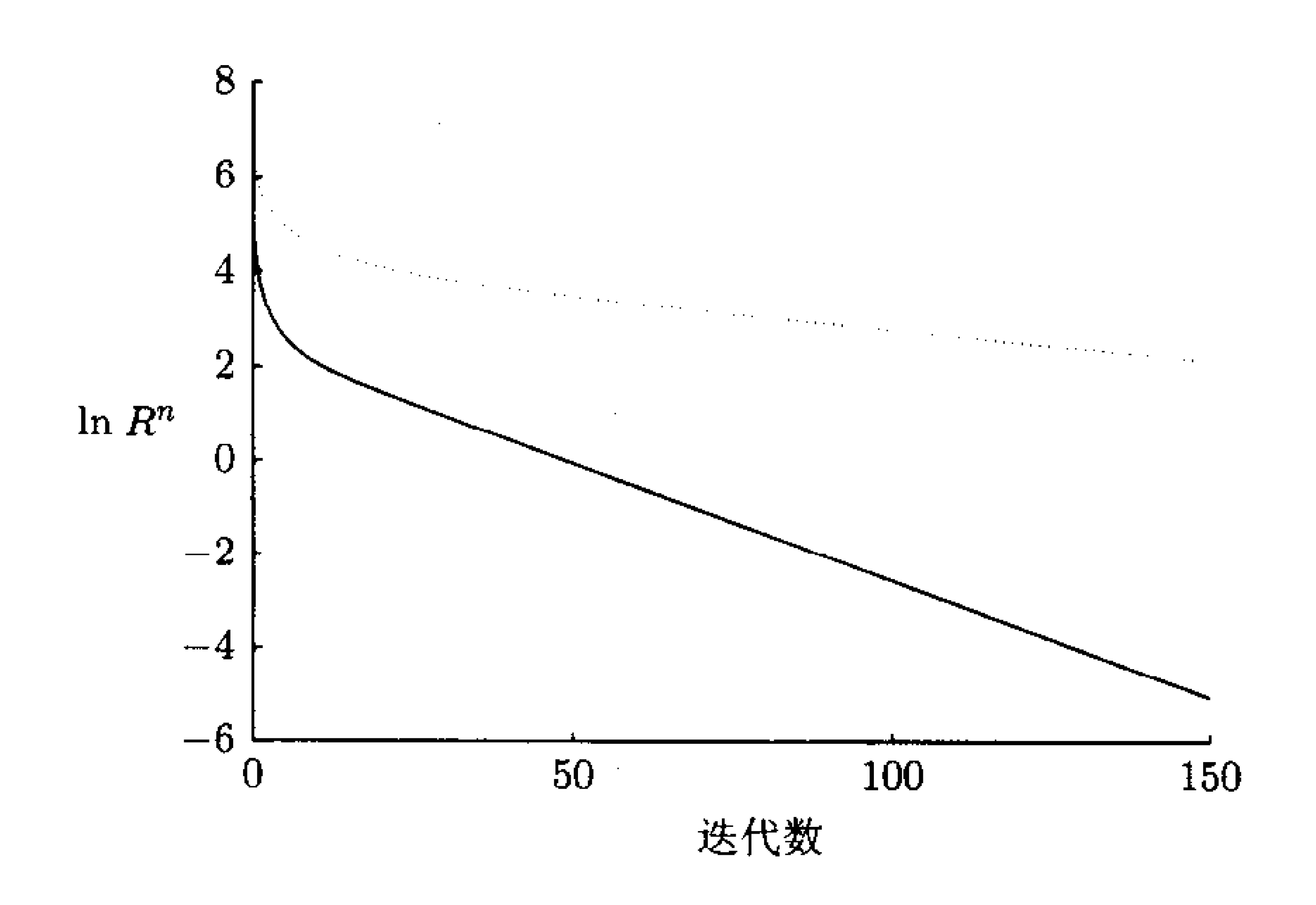


图 7-4 模型问题的高斯 - 赛德尔迭代: 下方曲线, $\Delta x = 1/10$; 上方曲线, $\Delta x = 1/20$

此图显示了两个重要特征: (i) 最初的少数几步迭代使残量迅速减小, 但之后的收敛速度则变得慢得多, (ii) $\Delta x = 1/20$ 的最终收敛速度要比 $\Delta x = 1/10$ 时慢了许多. 第二个观察正如 (7.38) 所预估. 稍后我们将在 7.8 节中再回顾这个问题.

图 7-5 画出了初始残量, 和经过 5 次, 20 次以及 50 次迭代后的残量分布的曲面图. 这些图的尺度不尽相同, 它们是通过重整后使其显出的最大误差的尺寸看上去差不多相同. 图示的重要性在于其形状而不是尺寸; 在非常参差不齐的最初的误差分布图之后, 它们迅速变得光滑起来, 甚至五次迭代之后, 随机的尖峰就大部分消失了.

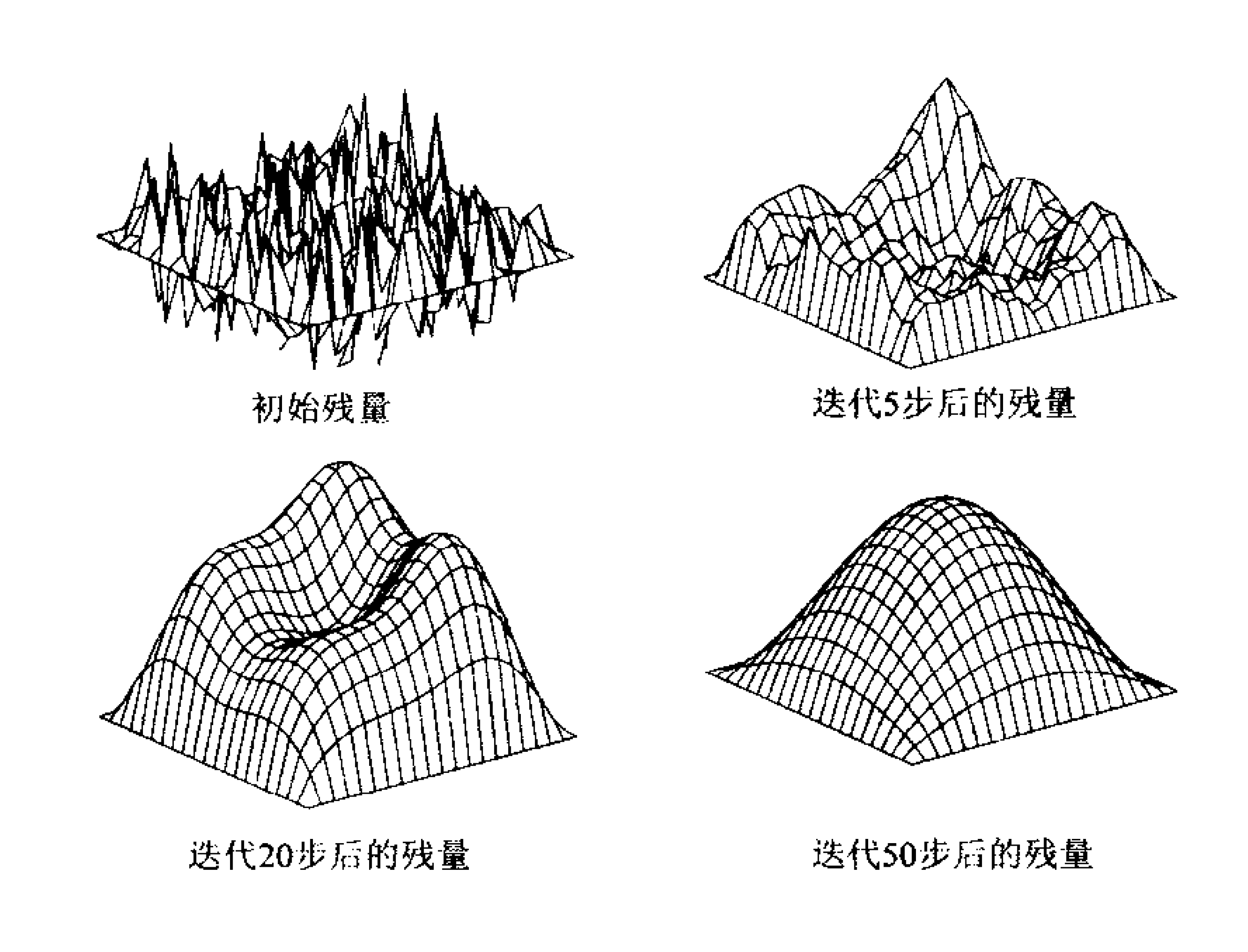


图 7-5 残量的曲面图

现在我们可以应用这些观察提出多重网格法的一个简单形式. 我们希望求解方程组

$$A\mathbf{x} = \mathbf{b}. \tag{7.52}$$

从初估计 $\mathbf{x}^{(0)}$ 开始，我们构造了一个序列 $\mathbf{x}^{(k)}$, $k = 1, 2, \ldots$, 相应的 *残量*(rcsidual) 为

$$\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}. \tag{7.53}$$

相应的 *误差*(error) $\mathbf{e}^{(k)}$ 则为向量 $\mathbf{e}^{(\mathbf{k})} = \mathbf{x} - \mathbf{x}^{(\mathbf{k})}$, 且有

$$A\mathbf{e}^{(\mathbf{k})} = \mathbf{r}^{(\mathbf{k})}. \tag{7.54}$$

下面假设残量 $\mathbf{r}^{(k)}$ 在某种可以精确定义的意义下是光滑的，那么向量 $\mathbf{r}^{(k)}$ 的各分量可以看作光滑函数 $r^k(x, y)$ 在网格点 (x_r, y_s) 处的值. 重要的是按比例正确地确定矩阵 A 的各行，从而代数方程组表示了微分方程的一个差分近似； A 的元素将包含一个因子 $1/(\Delta x)^2$. 那么代数方程 (7.54) 可以看作是微分方程

$$e_{xx} + e_{yy} = r \tag{7.55}$$

的一个有限差分近似. 假设函数 e 和 r 是光滑的，这样我们就可以在一个稍粗的网格上找到该微分方程的一个近似解. 通常，尽管不总是这样，可以通过间隔略去 x 和 y 方向的网格线而得到粗网格，这样得到的网格的尺寸是 Δx 和 Δy 的两倍. 重要的是这个新的有限差分问题所包含的未知数的个数是原离散问题的四分之一. 正如上面的分析已经证明的，这时迭代法的收敛通常会更快，并且每一步迭代所需要的工作量也更少.

我们已经勾勒了两重网格法的要点，它由四部分组成.

磨光 (smoothing) 首先选一种迭代法迭代少量几步磨光残量.

限制 (restriction) 计算残量并将它们传递到粗网格上；可以通过选用偶数标号的网格点处的值，或者通过若干相邻残量的加权平均完成.

粗网格校正 (coarse mesh correction) 在粗网格上求解方程组，将限制后的残量作为其右端向量.

延拓 (prolongation) 粗网格校正得到的解是定义在粗网格上的，通常可通过在 x 和 y 方向的线性中点插值将其延拓到细网格上. 结果是对向量 $\mathbf{e}^{(k)}$ 的近似，将其加到细网格的解 $\mathbf{x}^k$ 上便给出了原问题的解更好的近似.

后磨光 (post-smoothing) 在这一阶段通常再进行若干步磨光迭代.

两重网格法的核心步骤是在粗网格上求解方程组. 如果我们递归地进行这一步，即定义粗网格的新一层的放粗，我们用求解细网格上方程的方法来求解粗网格方程，那么这个方法即成为多重网格方法. 递归继续下去，每层网格的尺寸都是前一层的两倍，直到网格上未知量的个数小到求解代数方程直接法的效率与迭代法的效率一样高为止.

对于这四个阶段中的每一个，多重网格法都有非常广泛的选择；例如，7.1 节中介绍的任何一种迭代法都可以用于磨光，并且还有很多其他的选择．一个非常简单的例子是考虑对正方形区域上的泊松方程采用加权雅可比方法．

正如 SOR 推广了高斯 - 赛德尔迭代一样，加权雅可比方法推广了雅可比方法．它将 (7.8) 替换为

$$D\mathbf{x}^{(n+1)} = (1-\omega)D\mathbf{x}^{(n)} + \omega(\mathbf{b} + (L+U)\mathbf{x}^{(n)}).$$

其中 ω 是一个待定因子． 7.3 节的傅里叶分析确定了 (7.26) 的迭代矩阵的特征值，但是现在有

$$\mu = 1 - \omega(\sin^2 k_x\Delta x + \sin^2 k_y\Delta y).$$

光滑的特征值可以通过粗网格校正处理，而磨光过程只需要考虑 $k_x \geqslant \frac{1}{2}J\pi$ 或 $k_y \geqslant \frac{1}{2}J\pi$ 或两者兼有的傅里叶波型．此时最优选择是 $\omega = \frac{4}{5}$，它使得各相关傅里叶波型缩减超过 $\frac{3}{5}$．证明留作练习．一个重要的结果是这个磨光因子与网格尺寸无关．

我们现在将多重网格法应用到 Briggs, Henson 和 McCormick 用过的一个特殊的模型问题的求解上；这是一个单位正方形上的泊松方程，函数 f 的选取使得 $u(x,y) = (x^2-x^4)(y^4-y^2)$，它在正方形的边界上为零．我们用加权雅可比方法迭代 2 遍作为磨光子，另做两遍用来作后磨光；用全加权做限制，即用 9 个相邻细网格点上残量的加权平均计算粗网格残量．粗网格校正用一步相同的多重网格法；在最细的网格上，每一个方向都分为 $n = 2^p$ 个区间，最粗的网格只有一个内点，所以在这个网格上的求解是平凡的．初始近似处处取值为零．图 7-6 显示了数值计算的结果；它给出了每步迭代后的 $\ln(\|r^{(k)}\|)$．这里有五条曲线，分别对应 $\Delta x = 2^{-p}, p = 4,5,6,7,8$；各曲线接近平行，说明多重网格法的收敛速度与网格尺寸无关 (与图 7-4 比较)．事实上，除了在最初两步迭代中相对较小的 p 值的残量缩减得稍快些之外，这些曲线是相同的．

多重网格法需要从给定的细网格出发构造出粗网格．在我们的例子中，这样做是直接的，只涉及了偶数编号的网格线．如果我们在非矩形区域上构造三角形网格，如图 6-5，那么更方便的做法是首先定义最粗的网格，然后剖分每个三角形，将顶点取在每边的中点上，建立一列越来越细的网格．

在磨光阶段，我们可以在很广泛的迭代格式中选取其中的任何一个；其选择与问题的性质有关．例如，如果求解对流占优问题，如 (6.24)，其中 b 和 c 远大于 a，并且在区域的不同部分可能取不同的值、不同的符号，那么简单的加权雅可比磨光迭代就不是最高效的了．多重网格的文献中报道的众多研究成果有助于我们做出选择．

选定磨光格式后，有必要决定应该使用磨光子做几步迭代，一般只要一到两次，有时多些．在粗网格校正阶段，对粗网格递归应用多重网格法时，还有一个类似的决定需

要做出；这里可以将多重网格校正做多步迭代．只做一步迭代得到的方法称作 *V-循环*(V-cycle) 的方法，做两步迭代得到的是 *W-循环*(W-cycle)．两步以上的迭代则很少用到．

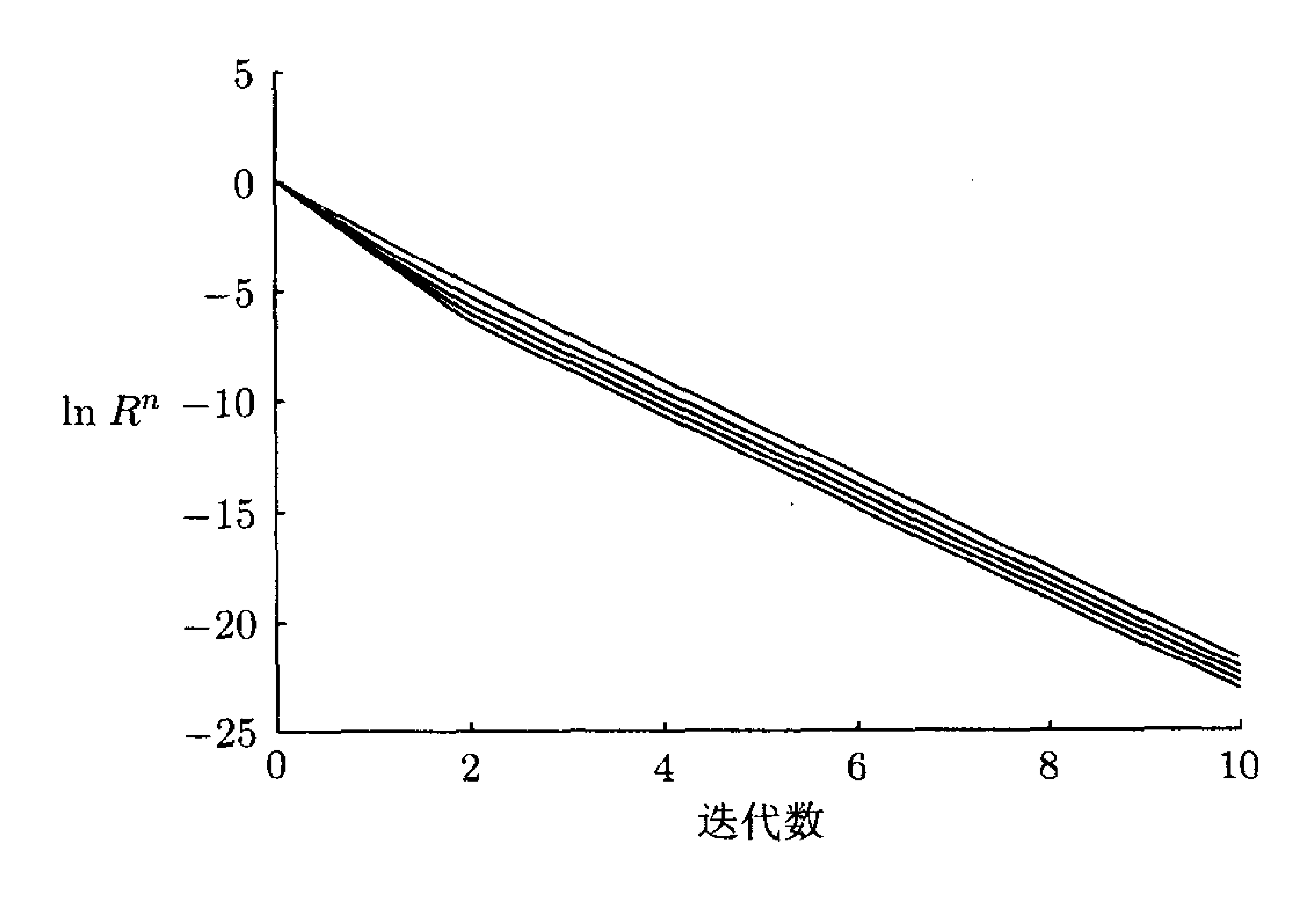


图 7-6 不同网格尺寸下多重网格的收敛性：从下方曲线开始： $\Delta x = 2^{-p}$, $p = 4, 5, 6, 7, 8$

多重网格法关键的重要性质是其总体收敛速度与网格尺寸无关．而如图 7-4 所示，当网格尺寸减小时，经典迭代法的收敛速度则会变慢．多重网格法的理论分析引入了 *工作量单元*(work unit) 的概念，它指的是在最细的网格上做一遍完整的多重网格迭代所用的运算量．工作量单元显然正比于网格点上未知数的个数．多重网格法要在少量的工作量单元内，通常 10 个左右，得到精确的解，这一点已经在非常广泛的问题中得以实现．

本章至今，我们只讨论了线性问题，但是，大部分实际问题当然都会涉及到非线性微分方程．多重网格法已推广至这些问题，且仍具有相同的效用．一个直接的方法是应用牛顿法求解整体非线性代数方程组．牛顿法的每步迭代所包含的大规模 *线性*(linear) 方程组则可以用我们介绍的多重网格法求解，这就涉及到一个双层迭代．应用更普遍的方法是所谓 *全逼近格式* (full approximation scheme), 本章末所列的参考书中对此有全面地论述．这是对我们所描述的多重网格格式的一种改进，它需要用某种非线性迭代作为磨光子，还需要在最粗的网格上求解一个小规模非线性方程组．认真选择这个方法的不同组成部分，仍然可以通过少数几个工作量单元求得精确的解．

7.7 共轭梯度法

上一节讨论的多重网格法利用了生成线性方程组的网格的结构．现在我们来继续考虑另一个利用了方程组的代数结构的求解大型线性代数方程组的高效迭代方法． Hestenes

和 Stiefel[1] 发明的共轭梯度法适用于由一大类具有对称正定结构的问题所导出的方程组.

我们从求解方程组 $A\mathbf{x}=\mathbf{b}$ 入手，这里矩阵 A 是对称正定的；在实际问题中，它当然还是稀疏的. 这种方法基于这样的观察，即函数

$$\phi(\mathbf{x})=\frac{1}{2}\mathbf{x}^TA\mathbf{x}-\mathbf{b}^T\mathbf{x} \tag{7.56}$$

唯一的最小值点 $\mathbf{x}^*$ 是方程组 $A\mathbf{x}=\mathbf{b}$ 的解； $\phi(\cdot)$ 的相应的值是 $\phi(\mathbf{x}^*)$.[2] 共轭梯度法要构造一列向量 $\mathbf{x}^{(k)}$, 使得 $\phi(\mathbf{x}^{(k)})$ 构成了一个单调递减的序列，且收敛到最小值 $\phi(\mathbf{x}^*)$.[3]

假设我们对解有一个估计 $\mathbf{x}^{(k)}$, 并且有办法选择一个由 $\mathbf{p}^{(k)}$ 定义的*搜索方向* (search direction), 则序列中的下一个向量可以选择为

$$\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}+\alpha_k\mathbf{p}^{(k)}, \tag{7.57}$$

其中标量 α_k 的选择使 $\phi(\mathbf{x}^{(k)}+\alpha_k\mathbf{p}^{(k)})$ [4]作为 α_k 的函数取得最小值. 通过微分容易证明所需要的最小值点取在

$$\alpha_k=\mathbf{p}^{(k)T}\mathbf{r}^{(k)}/\mathbf{p}^{(k)T}A\mathbf{p}^{(k)}, \tag{7.58}$$

其中 $\mathbf{r}^{(k)}$ 是残量向量

$$\mathbf{r}^{(k)}=\mathbf{b}-A\mathbf{x}^{(k)}. \tag{7.59}$$

选择最佳的搜索方向 $\mathbf{p}^{(k)}$ 显然是重要的. 共轭梯度法的基本性质是，它构造了一列向量 $\mathbf{p}^{(k)}$ 使得当我们由 (7.58) 确定 α_k, 并继而由 (7.57) 确定 $\mathbf{x}^{(k+1)}$ 时，新向量 $\mathbf{x}^{(k+1)}$ 给出的 $\phi(\mathbf{x})$ 的最小值点不仅是搜索方向上的，而且是由搜索方向 $\mathbf{p}^{(0)},\cdots,\mathbf{p}^{(k)}$ 张成的整个子空间上的. 下面我们将看到这个关键性质来自于各方向是“A- 共轭” (A-conjugate) 的，即 $\mathbf{p}^{(k)T}A\mathbf{p}^{(j)}=0,\ j<k$.

这种方法的基本形式是收敛的，但是对于我们考虑的那类问题来说，收敛速度通常相当令人失望. 应用*预条件* (preconditioning) 可以改变这种局面. 这种思想的出发点是对任何一个非奇异矩阵 P, $A\mathbf{x}=\mathbf{b}$ 的解与 $PA\mathbf{x}=P\mathbf{b}$ 的解相同；适当地选择 P 可能会使新的方程组容易求解得多. 矩阵 PA 通常不是对称的，所以我们将这个想法再推进一步，选择非奇异矩阵 C, 它还具有我们下面会讨论到其他一些性质，并将方程组写成

$$\left(C^{-1}A(C^{-1})^T\right)C^T\mathbf{x}=C^{-1}\mathbf{b}. \tag{7.60}$$

这个方程组的矩阵是对称的；在从此方程组中解出向量 $C^T\mathbf{x}$ 后，我们还需要还原所要求的向量 $\mathbf{x}$. 不过，记 $M=C^TC$ 会使过程简化，现在就取它作为预条件子 (preconditioner)；

[1] Hestenes, M.R. and Stiefel, E.(1952). Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.* **49**, 409-36.

[2] 原书这里是： the correspongding value of $\phi(\mathbf{x})$ is zero.

[3] 原书这里是： converging to zero.

[4] 原书为 $\phi(\mathbf{x}^{(k)})$

这时预条件共轭梯度算法有如下形式：

首先令 $\mathbf{x}^{(0)}=\mathbf{0}$, $\mathbf{r}^{(0)}=\mathbf{b}$. 求解 $M\mathbf{z}^{(0)}=\mathbf{b}$, 并令 $\beta_1=0$, $\mathbf{p}^{(1)}=\mathbf{z}^{(0)}$. 然后对 $k=1,2,\ldots$ 执行下列步骤：

$$\begin{aligned}
\alpha_k &= \mathbf{z}^{(k-1)T}\mathbf{r}^{(k-1)}/\mathbf{p}^{(k)T}A\mathbf{p}^{(k)},\\
\mathbf{x}^{(k)} &= \mathbf{x}^{(k-1)}+\alpha_k\mathbf{p}^{(k)},\\
\mathbf{r}^{(k)} &= \mathbf{r}^{(k-1)}-\alpha_k A\mathbf{p}^{(k)},\ \text{并求解 } M\mathbf{z}^{(k)}=\mathbf{r}^{(k)},\\
\beta_{k+1} &= \mathbf{z}^{(k)T}\mathbf{r}^{(k)}/\mathbf{z}^{(k-1)T}\mathbf{r}^{(k-1)},\\
\mathbf{p}^{(k+1)} &= \mathbf{z}^{(k)}+\beta_{k+1}\mathbf{p}^{(k)}.
\end{aligned} \tag{7.61}$$

注意计算中没有用到矩阵 $C^{-1}A(C^{-1})^T$, 只用了矩阵 M 和原始矩阵 A. 可以证明残量向量 $\mathbf{r}$ 和搜索方向 $\mathbf{p}$ 满足关系：对 $i\neq j$ 有

$$\begin{aligned}
\mathbf{r}^{(j)T}M^{-1}\mathbf{r}^{(i)} &= 0,\\
\mathbf{p}^{(j)T}(C^{-1}A(C^{-1})^T)\mathbf{p}^{(i)} &= 0.
\end{aligned} \tag{7.62}$$

这构成了该算法收敛的基础.

预条件子的选择对方法的成功非常重要. 这里有两个明显的极端情形：取 $M=I$, I 是单位矩阵，对应于不用预条件的基本方法，该方法收敛得太慢；另一方面，如果我们选择 $M=A$, 容易看出，一步之后，求得精确解，过程终止. 但这并非实用的选择，因为我们必须要能够容易地求解方程组 $M\mathbf{z}=\mathbf{r}$. 为构造实用的方法，我们需要一个在某种意义上接近 A, 但又使 $\mathbf{z}$ 容易解出的矩阵 M.

利用不完全 Cholesky 分解构造 M 是一种频繁使用的方法，由此导出的方法通常称为 ICCG 算法. Cholesky 分解算法构造一个下三角矩阵 L 使 $A=LL^T$. 与其他的分解方法类似，L 通常是一个非常满的矩阵，甚至在 A 非常稀疏时也不例外. 在 A_{ij} 不为零的位置，不完全 Cholesky 分解利用相同的公式计算 L_{ij} 的元素，但是在 $A_{ij}=0$ 处，取 $L_{ij}=0$. 结果是 L 与 A 具有相同的稀疏结构，并且 LL^T 相当接近于 A. 与其他的分解过程类似，并不需要构造出预条件矩阵 M 本身. 构造出 L 后，可以通过求解两个三角型方程组得到向量 $\mathbf{z}^{(k)}$, 即

$$L\eta=\mathbf{r}^{(k)},\qquad L^T\mathbf{z}^{(k)}=\eta. \tag{7.63}$$

由于 L 的稀疏性，整个计算都非常高效.

我们刚看到对 ICCG 方法，没有必要显式构造出矩阵 M, 而且对更一般地来说也是如此. 我们所需要的只是能够找到一个接近 $A\mathbf{z}=\mathbf{r}$ 的解的向量 $\mathbf{z}$, 并且对某个对称正定矩阵 M, 该向量精确满足 $M\mathbf{z}=\mathbf{r}$. 这就意味着我们可以考虑众多迭代法中的任一个. 譬如我们可以用求解方程组 $A\mathbf{z}=\mathbf{r}$ 的雅可比方法迭代若干步，因为这时的迭代矩阵是对称的.

SOR 方法的迭代矩阵不是对称的，尽管仍然可能存在对称矩阵 M, 使得迭代的结果是其精确解. 在实际问题中，各种预条件子的效率必须通过实验核定. 在很多情况下，我们发现 SOR 迭代法是很好的预条件子. 不完全 Cholesky 算法似乎对一大类问题都有效，但是它没有用到导出方程组的偏微分方程的特殊性质. 得到研究和推荐的方法种类繁多，比如利用 7.5 节中所讨论的多种分块方法，以及相关的区域分解的思想. 预条件子的选取对收敛速度有决定性的影响，而且难以像处理多重网格法那样给出简单的傅里叶分析.

我们所介绍的共轭梯度法要求矩阵 A 是对称正定的. 将这种思想向非对称和 (或) 非正定的矩阵的推广工作已经有一些结果，但是迄今为止，它们的实际应用效果还不像原型那么令人满意；当前这方面的研究很活跃. 另一种不同但相关联的方法是直接极小化残量的平方和. 这种方法可以应用到任意矩阵 A: 最广泛使用的算法是 Saad 和 Schultz[1] 提出的 GMRES. 类似于共轭梯度法，它构造了一个搜索方向序列 $\mathbf{p}^{(k)}$, 但是构造 $\mathbf{p}^{(k+1)}$ 时用到了前面所有的向量 $\mathbf{p}^{(1)},\ldots,\mathbf{p}^{(k)}$. 注意，共轭梯度法仅从 $\mathbf{p}^{(k-1)}$ 和 $\mathbf{p}^{(k)}$ 出发构造向量 $\mathbf{p}^{(k+1)}$, 这节省了计算量和存储量. 由于这两种方法都利用了初始向量的相继的累次幂 $A\mathbf{z}^{(0)}, A^2\mathbf{z}^{(0)}, A^3\mathbf{z}^{(0)},\cdots$, 因此它们都属于一类更一般的 Krylov 子空间法 (Krylov subspace method).

7.8 数值例子：几个对比

作为本章的结尾，我们在图 7-7 中画出将三种迭代方法应用于图 7-4 中用过的问题

$$(a(x,y)u_x)_x + (a(x,y)u_y)_y + f(x,y) = 0, \qquad 0 < x < 1, 0 < y < 1, \tag{7.64}$$

所得到的数值结果，其中 $a(x,y) = (x + 2y^2 + 1)$, 解是 $u(x,y) = x(1-x)y(1-y)$. 数值实验中取 $J = 40$, 所以代数方程组的阶是 1521.

我们首先采用 7.2 节中给出的对称 SOR 方法. 由于系数 $a(x,y)$ 不是常数，问题比 7.6 节中的更一般些，并且我们也无法利用傅里叶分析来寻找最佳松弛因子 ω. 我们在一定范围试验了多个值，发现在实验中 $\omega = 1.85$ 给出最快收敛速度；这实际上接近于由 (7.42) 预估的最优值. 在图 7-7 中，顶部的曲线表示相应迭代法的收敛速度；收敛正如预计的一样慢， 50 步迭代后残量缩小为 1/55.

底部曲线表示应用多重网格的结果，其中用高斯 - 赛德尔迭代作为磨光子，并有两步放粗. 仅在 8 步迭代后，它就给出了精确到六位 10 进制小数的解. 中间的曲线采用预条件共轭梯度法，不完全 Cholesky 分解用作预条件子，在第一步迭代后，它显示的收敛速度非常类似于多重网格.

[1] Saad, Y. and Schultz, M.H.(1986). A generalised minimal residual algorithm for solving nonsymmetric linear systems. *J. Sci. Stat. Comput.* **7**, 856-69.

做这个计算的目的并不在于比较多重网格与 ICCG, 而是要说明它们两者都要比经典的 SOR 迭代有效得多. 当然了, 它们也都更复杂些, 所以每步迭代也要比一步 SOR 迭代花费更多的工作量. 不过即使将这些计算在内, 它们二者也仍要比 SOR 快若干量级.

多重网格和共轭梯度这两种方法应该看作是互为补充的, 而不是相互竞争的. Elman, Silvester 和 Wathen(2004) 中公布的, 关于定常不可压流的 Navier-Stokes 方程的最近的研究中采用了共轭梯度法, 并用多重网格迭代做预条件子. 结果显示, 在少量的工作量单元内即收敛, 而且收敛所需的步数几乎与网格节点数无关.

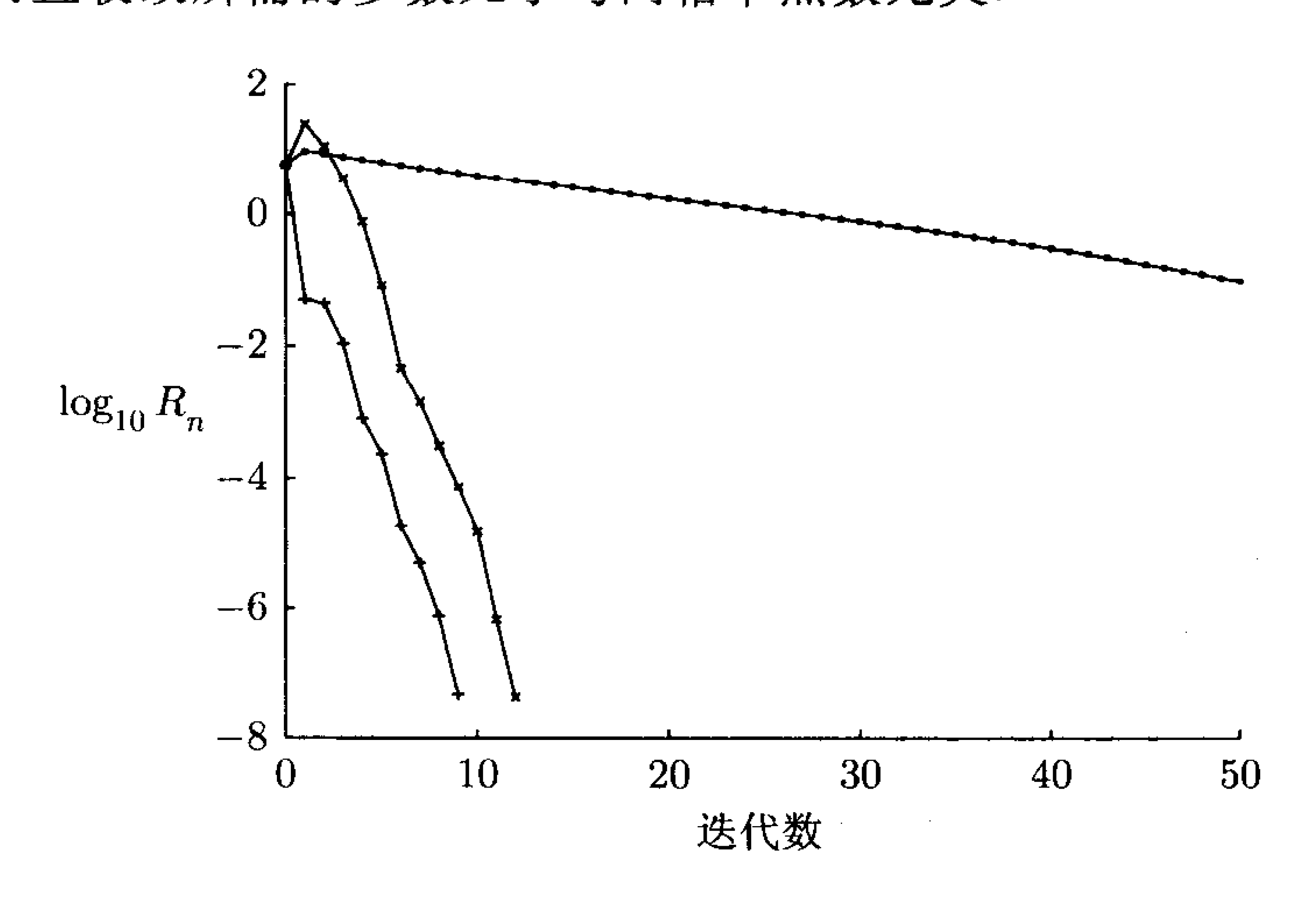


图 7-7 迭代法用到模型问题 (7.64) 的比较

顶部曲线: SOR 取 $\omega = 1.85$

中部曲线: 不完全 Cholesky 共轭梯度

底部曲线: 多重网格

文献注记与推荐读物

Varga(1974) 给出了线性方程组迭代方法经典理论的全面论述, 稍早些的 Young(1971) 也有类似的论述. Duff, Erisman 和 Reid(1986) 全面涵盖了充分利用矩阵稀疏性的直接方法. Goluf 与 Van Loan(1996) 和 Hackbusch (1994) 详尽地论述了本章中所介绍的许多论题.

Briggs, Henson 和 McCormick(2000) 的多重网格指南是一本有用的入门书, 而 Wesseling(1992) 和 Hackbusch (1985) 则详尽分析了多重网格方法.

共轭梯度法在很多书中都有所论及，例如 Golub 和 Van Loan(1996). Van der Vorst(2003) 和 Elman, Silvester 与 Wathen(2004) 给出了更多细节. 这是非常活跃的研究领域，这些书可以帮助读者了解最近的有关文献.

习　题

7.1　设 2×2 阶矩阵 A 是对称正定的；求证关于 A 的雅可比迭代法收敛.

7.2　设矩阵 A 是对称正定的，将其写成 $A=L+D+L^T$ 的形式，其中 L 是严格下三角的. 假设 λ 和 $\mathbf{z}$ 是 A 的高斯 - 赛德尔迭代矩阵的特征值和特征向量，满足 $\bar{\mathbf{z}}^T D\mathbf{z}=1$. 求证

$$\lambda=-\frac{\bar{\mathbf{z}}^T L^T\mathbf{z}}{1+\bar{\mathbf{z}}^T L\mathbf{z}}.$$

并证明 $2p+1>0$ 和 $|\lambda|<1$, 这里 p 是 $\bar{\mathbf{z}}^T L\mathbf{z}$ 的实部. 推演出高斯 - 赛德尔方法对任意对称正定矩阵都收敛.

7.3　$N\times N$ 阶矩阵 E 的所有元素都等于 1. 求证 E 的一个特征值是 N, 并且其他的特征值都是 0. 构造一个矩阵 $A=I+kE$, 这里 k 是一个待定常数，满足 A 对称正定，但是雅可比迭代法一般不收敛. 解释为什么这个结果与练习 7.1 不相矛盾.

7.4　用标准五点差分格式求解建立在单位正方形上赋狄利克雷边界条件的泊松方程，区间为 $\Delta x=\Delta y=\frac{1}{3}$, 得方程组 $A\mathbf{u}=\mathbf{b}$. 矩阵 A 是 4×4 的. 按照不同顺序给网格点标号，得到 24 个这样的矩阵. 考虑问题的对称性，则只有 3 个不同的矩阵，每个各出现 8 次. 确定雅可比迭代矩阵和 3 个高斯 - 赛德尔迭代矩阵的特征值. 验证对于其中的两个矩阵，其高斯 - 赛德尔矩阵的特征值分别是 0 和其雅可比矩阵的特征值的平方.

7.5　为在单位正方形上求解 $\nabla^2 u+f=0$, 在 $x=0$ 给定了诺伊曼条件，在另外三边给定了狄利克雷条件. 我们在对应 $x=-\Delta x$ 处额外加一列点，并用中心差分格式表示法向导数，以此近似诺伊曼条件. 得到的线性方程组用雅可比迭代法求解. 证明误差可以用傅里叶波型展开为

$$e_{r,s}^{(n)}=\sum_{k_x,k_y} a^{(n)}(k_x,k_y)\cos k_x x_r \sin k_y y_s,$$

其中 $k_y=\pi,\ 2\pi,\ \ldots,\ (J-1)\pi,\ \ k_x=\frac{1}{2}\pi,\ \frac{3}{2}\pi,\ \ldots,\ (J-\frac{1}{2})\pi$. 推演出雅可比迭代矩阵的特征值是

$$\mu_J(k_x,k_y)=\frac{1}{2}(\cos k_x\Delta x+\cos k_y\Delta y),$$

并且对于 $\Delta x=\Delta y$ 的正方形网格，谱半径约为 $1-\frac{5}{16}(\pi\Delta x)^2$.

将此分析扩展至诺伊曼条件给定在正方形的两条边上问题，然后是三条边上的问题. 若在边界所有点处给定诺伊曼条件，会是什么样子？

7.6 三角形区域 $x \geqslant 0, y \geqslant 0, x+y \leqslant 1$ 上的拉普拉斯方程 $u_{xx}+u_{yy}=0$ 用尺寸为 $\Delta x=\Delta y=1/N$ 的正方形网格上的中心差分格式逼近. 用雅可比迭代法求解得到线性方程组. 证明分量为

$$w_{r,s}=\sin\left(\frac{pr\pi}{N}\right)\sin\left(\frac{qs\pi}{N}\right)\pm\sin\left(\frac{qr\pi}{N}\right)\sin\left(\frac{ps\pi}{N}\right)$$

的向量是迭代矩阵的特征向量. 从而确定迭代矩阵的特征值, 以及渐近收敛速度.

7.7 在尺寸 $\Delta x=\Delta y$ 的一致正方形网格上用三项差分格式

$$\frac{U_{r,s}^{n+1}-U_{r,s}^{n-1}}{2\Delta t}=\frac{U_{r,s+1}^{n}+U_{r,s-1}^{n}+U_{r+1,s}^{n}+U_{r-1,s}^{n}-2U_{r,s}^{n+1}-2U_{r,s}^{n-1}}{(\Delta x)^2}+f_{r,s}$$

求解方程 $u_t=u_{xx}+u_{yy}+f(x,y)$. 确定其冯诺伊曼稳定性条件.

说明这个格式可以如何用作求解方程

$$\delta_x^2U_{r,s}+\delta_y^2U_{r,s}+(\Delta x)^2f_{r,s}=0.$$

的迭代法. 证明取 $\Delta t=\frac{1}{4}(\Delta x)^2$ 得到等同于雅可比方法的收敛速度, 但若选

$$\Delta t=\frac{1}{4}(\Delta x)^2/\sin(\pi\Delta x)$$

则使收敛显著增快.

7.8 证明定义为

$$D\mathbf{x}^{n+1}=(1-\omega)D\mathbf{x}^n+\omega(\mathbf{b}+(L+U)\mathbf{x}^n).$$

的加权雅可比方法的迭代矩阵是

$$G=[(1-\omega)I+\omega D^{-1}(L+U)].$$

证明对于 6.1 节中的问题, 该矩阵的特征值是

$$\mu_{r,s}=1-\omega(\sin^2\frac{r\pi}{2J}+\sin^2\frac{s\pi}{2J}),\qquad r,s=1,2,\ldots,J-1.$$

定义对应于 $2r\leqslant J$ 和 $2s\leqslant J$ 的特征向量是光滑特征向量, 验证相应于非光滑特征向量的特征值在 $1-\frac{1}{2}\omega$ 到 $1-2\omega$ 的范围内. 证明选择 $\omega=2/5$ 时磨光子的效果最佳, 这样选取后, 误差的每个非光滑分量都会缩减至少 3/5 倍.

7.9 将上题中的问题换为单位立方体上求解赋狄利克雷边界条件的泊松方程的三维问题. 找出 ω 的最佳取法, 并证明这时加权雅可比迭代将误差的每个非光滑分量缩减至少 5/7 倍.

7.10 在由 (7.61) 定义的, 相应于取 $M=I$ 的, 共轭梯度迭代中, 归纳证明对每个正整数 n, 向量 $\mathbf{x}^{(n)}$, $\mathbf{p}^{(n-1)}$ 和 $\mathbf{r}^{(n-1)}$ 都属于由向量 $\mathbf{b}, A\mathbf{b},\ldots,A^{n-1}\mathbf{b}$ 张成的 Krylov 空间 K^n 中.

现假设 $N\times N$ 矩阵 A 只有 k 个不同的特征值, 这里 $k\leqslant N$. 证明空间 K^n 的维数不超过 k, 并推导出在至多 $k+1$ 步后, 迭代终止并得到 $A\mathbf{x}=\mathbf{b}$ 的精确解.

7.11 (7.61) 给出的共轭梯度算法始于初始近似 $\mathbf{x}^{(0)}=0$. 假设已知解的一个近似 x^*, 如何修改此方法来利用这条信息?

其他参考文献

[下列文献由各章的文献注记与推荐读物中推荐的文献汇编而成，正文其他地方提到的文献在脚注中列出.]

Ames,W.F. (1995). *Nonlinear Partial Differential Equations in Engineering*, Vol. I. New York,Academic Press.

(1972). *Nonlinear Partial Differential Equations in Engineering*,Vol. II. New York, Academic Press.

(1992). *Numerical Methods for Partial Differential Equations*, 3rd edn. Boston, Academic Press.

Brenner, S. and Scott, L. R. (2002). *The Mathematical Theory of Finite Element Methods*, Second Edition. New York,Springer.

Briggs, W. L., Henson, V. E. and McCormick, S. F.(2000). *A Multigrid Tutorial*, Second edition. SIAM.

Carrier, G. F. and Pearson, C. E. (1976). *Partial Differential Equations.* New York, Academic Press.

Ciarlet, P. G. (1978). *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam.

Collatz, L. O. (1996). *The Numerical Treatment of Differential Equations.* Berlin, Springer.

Courant, R. and Hilbert , D. (1962). *Methods of Mathematical Physics, Vol2: Partial Differential Equations.* NewYork, Wiley-Interscience.

Crank, J. (1975). *The Mathematics of Diffusion*, 2nd edn. Oxford, Clarendon Press.

Duff, I. S., Erisman, A. M. and Reid, J. K.(1986). *Direct Methods for Sparse Matrices.* Oxford, Clarendon Press.

Elman, H. C., Silvester, D. J. and Wathen, A. J. (2004). *Finite Elements and Fast Iterative Solvers.* Oxford, Oxford University Press.

Evans, L. C.(1998). *Partial Differential Equations,* Graduate Studies in Mathematics. Providence, Rhode Island, American Mathematical Society.

Godunov, S. K. and Ryabenkii, V. S. (1964). *The Theory of Difference Schems-An Introduction.* North Holland, Amsterdam.

Golub, G. H. and Van Loan, C. F.(1996). *Matrix Computations,* 3rd edn. Baltimore, Johns Hopkins University Press.

Gottlieb, S., Shu, C. W. and Tadmor, E. (2001). Strong stabilitypreserving high-order time Discretisation methods. *SIAM Rev.* **43**, 89-112.

Hackbusch, W. (1985). *Multigrid methods and applications.* Berlin, Springer-Verlag.
(1994).*Iterative solution of large sparse linear systems of equations.* Berlin, Springer-Verlag.

Hairer, E., Lubich, C. and Wanner, G. (2002). *Geometric Numerical Integration.* Berlin, Springer- Verlag.

Kreisss, H. O. and Lorenz, J. (1989). *Initial-Boundary Value Problems and the Navier-Stokes Equations.* Academic Press, San Diego.

Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian Dynamics.* Cambridge, Cambridge University Press.

LeVeque, R. J. (1992). *Numerical Methods for Conservation Laws,* Lectures in Mathematics ETH Zurich, 2nd edn. Basel, Birkhauser Verlag
(2002).*Finite Volume Methods for Hyperbolic Problems.* Cambridge, Cambridge University Press.

Lighthill, M. J. (1978) . *Waves in Fluids.* Cambridge, Cambridge University Press.

Mitchell, A. R. and Griffiths, D. F. (1980). *The Finite Difference Method in Partial Differential Equations.* New York, Wiley-Interscience.

Morton, K. W. (1996). *Numercial Solution of Convection-Diffusion Problems.* London, Chapman & Hall.

Protter, M. H. and Weinberger, H. F. (1967). *Maximum Principles in Differential Equations.* Englewood Cliffs, NJ, Prentice-Hall.

Reich, S. (1999). Backward error analysis for numerical integration. *SIAMJ. Numer. Anal.* **36**, 1549-1570.

Richtmyer, R. D. and Morton, K. W. (1967). *Difference Methods for Initial Value Problems*, 2nd edn. New York, Wiley-Interscience. Reprinted (1994), New York, Kreiger.

Roos, H. G. , Stynes, M. and Tobiska, L. (1996). *Numerical Methods for Singularly Perturbed Differential Equations*. Berlin, Springer.

Shokin, Y. I. (1983). *The Method of Differential Approximation.* New York-Berlin, Springer.

Smoller, J. (1983). *Shock Waves and Reaction-Diffusion Equations.* New York, Springer-Verlag.

Strang, G. and Fix, G. (1973). *An Analysis of the Finite Element Method.* New York, Prentice-Hall.

Süli, E. and Mayers, D. F. (2003). *An Introduction to Numerical Analysis.* Cambridge, Cambridge University Press.

van der Vorst, H. A. (2003). *Iterative Krylov Methods for Large Linear Systems.* Cambridge, Cambridge University Press.

Varga, R. S. (1974). *Matrix Iterative Analysis.* Englewood Cliffs, NJ, Prentice-Hall.

Wesseling, P. (1992). *An Introduction to Multigrid Methods.* Chichester, John Wiley.

Whitham, G. B. (1974). *Linear and Nonlinear Waves.* New York, Wiley-Interscience.

Yanenko, N. N. (1971). *The Method of Fractional Steps.* Berlin, Springer-Verlag.

Young, D.M . (1971). *Iterative Solutions of Large Linear Systems.* New York, Academic Press.